Behavioral/Cognitive

Linear Integration of Sensory Evidence over Space and Time Underlies Face Categorization

Gouki Okazawa,[1] Long Sha,[1] and Roozbeh Kiani[1,2,3]

[1]Center for Neural Science, New York University, New York, New York 10003, [2]Neuroscience Institute, New York University Langone Medical Center, New York, New York 10016, and [3]Department of Psychology, New York University, New York, New York 10003

Visual object recognition relies on elaborate sensory processes that transform retinal inputs to object representations, but it also requires decision-making processes that read out object representations and function over prolonged time scales. The computational properties of these decision-making processes remain underexplored for object recognition. Here, we study these computations by developing a stochastic multifeature face categorization task. Using quantitative models and tight control of spatiotemporal visual information, we demonstrate that human subjects (five males, eight females) categorize faces through an integration process that first linearly adds the evidence conferred by task-relevant features over space to create aggregated momentary evidence and then linearly integrates it over time with minimum information loss. Discrimination of stimuli along different category boundaries (e.g., identity or expression of a face) is implemented by adjusting feature weights of spatial integration. This linear but flexible integration process over space and time bridges past studies on simple perceptual decisions to complex object recognition behavior.

Key words: **face recognition; flexible decision making; linear spatiotemporal integration; feature combination; bounded accumulation of evidence; reverse correlation; psychophysics**

Significance Statement

Although simple perceptual decision-making such as discrimination of random dot motion has been successfully explained as accumulation of sensory evidence, we lack rigorous experimental paradigms to study the mechanisms underlying complex perceptual decision-making such as discrimination of naturalistic faces. We develop a stochastic multifeature face categorization task as a systematic approach to quantify the properties and potential limitations of the decision-making processes during object recognition. We show that human face categorization could be modeled as a linear integration of sensory evidence over space and time. Our framework to study object recognition as a spatiotemporal integration process is broadly applicable to other object categories and bridges past studies of object recognition and perceptual decision-making.

Introduction

Accurate and fast discrimination of visual objects is essential to guide our behavior in complex and dynamic environments. Previous studies largely focused on the elaborate sensory mechanisms that transform visual inputs to object-selective neural responses in the inferior temporal cortex of the primate brain through a set of representational changes along the ventral visual pathway (Riesenhuber and Poggio, 1999; DiCarlo and Cox, 2007; Freiwald and Tsao, 2010; Yamins et al., 2014). However, goal-directed behavior also requires decision-making processes that can flexibly read out sensory representations and guide actions based on them as well as information about the environment, behavioral goals, and expected costs and gains. Such processes have been extensively examined using simplified sensory stimuli that vary along a single dimension, for example, the direction of moving dots changing from left to right (Palmer et al., 2005). For those stimuli, subjects' behavior could be successfully accounted for by flexible mechanisms that accumulate sensory evidence and combine it with task-relevant information (Ratcliff and Rouder, 1998; Gold and Shadlen, 2007). However, more complex visual decisions based on stimuli defined by multiple features, such as object images, remain underexplored, although the need for such tests is gaining significance, and important steps are being taken in this direction (Heekeren et al., 2004; Philiastides and Sajda, 2006; Philiastides et al., 2014; Zhan et al., 2019).

Received Dec. 4, 2020; revised July 8, 2021; accepted July 21, 2021.

Author contributions: G.O. and R.K. designed research; G.O. and L.S. performed research; G.O. analyzed data; G.O. and R.K. wrote the paper.

This work was supported by the Simons Collaboration on the Global Brain (Grant 542997), McKnight Scholar Award, Pew Scholars Program in the Biomedical Sciences Award, and National Institute of Mental Health (Grant R01 MH109180-01). G.O. was supported by postdoctoral fellowships from the Charles H. Revson Foundation and the Japan Society for the Promotion of Science. We thank Stanley J. Komban, Michael L. Waskom, and Koosha Khalvati for discussions.

The authors declare no competing financial interests.

Correspondence should be addressed to Roozbeh Kiani at roozbeh@nyu.edu.

https://doi.org/10.1523/JNEUROSCI.3055-20.2021

Here, we apply the quantitative approach developed for studying simple perceptual decisions to investigate face recognition. We focus on face recognition because it is by far the most extensively studied among the subdomains of object vision (Kanwisher and Yovel, 2006; Tsao and Livingstone, 2008; Barraclough and Perrett, 2011; Rossion, 2014; Perrodin et al., 2015). Face stimuli are also convenient to use because they allow quantitative manipulation of sensory information pivotal for mechanistic characterization of the decision-making process (Waskom et al., 2019); images can be decomposed into local spatial parts (e.g., eyes, nose, mouth) and can be morphed between two instances (e.g., faces of two individuals) to create a parametric stimulus set. At the same time, human face perception is highly elaborate and embodies the central challenge of object recognition that must distinguish different identities from complex visual appearances (Tsao and Livingstone, 2008).

To quantitatively characterize the decision-making process, we investigate face recognition as a process of combining sensory evidence over both space and time. Faces are thought to be processed holistically (Maurer et al., 2002; Richler et al., 2012); breaking the configuration of facial images significantly affects face perception, indicating spatial interactions across facial parts. However, computational properties of the spatial integration remain elusive (Richler et al., 2012). One may consider that holistic recognition arises from nonlinear integration of facial features (Shen and Palmeri, 2015), but linear integration may also suffice to account for holistic effects (Gold et al., 2012). Furthermore, humans flexibly use different facial parts to categorize faces according to their behavioral needs (e.g., discrimination of identity vs expression; Schyns et al., 2002, 2007), but the underlying mechanisms of this flexibility also remain underexplored.

In addition to spatial properties, face and object recognition also include rich temporal dynamics. Although object identification and categorization are usually fast, reaction times (RTs) are often hundreds of milliseconds longer (Gauthier et al., 1998; Kampf et al., 2002; Ramon et al., 2011; Carlson et al., 2014; Witthoft et al., 2018) than the time required for a feedforward sweep along the ventral visual pathway (Thorpe et al., 1996; Hung et al., 2005). Furthermore, recognition performance follows a speed-accuracy trade-off, where additional time improves accuracy (Thorpe et al., 1996; Gauthier et al., 1997). Together, these observations suggest that the decision-making process in face and object recognition is not instantaneous but unfolds over time (Heitz and Schall, 2012; Hanks et al., 2014). However, the computational properties have scarcely been characterized.

Using our novel face categorization tasks that tightly control spatiotemporal sensory information (Okazawa et al., 2018, 2021), we show that human subjects categorize faces by linearly integrating visual information over space and time. Spatial features are weighted nonuniformly and integrated largely linearly to form momentary evidence, which is then accumulated over time to generate a decision variable that guides the behavior. The temporal accumulation is also linear, and the time constant is quite long, preventing significant loss of information (or leak) during the decision-making process. Between identity and expression categorizations, the weighting for spatial integration flexibly changes to accommodate task demands. Together, we offer a novel framework to study face recognition as a spatiotemporal integration process, which unifies two rich veins of visual research, namely, object recognition and perceptual decision-making.

Materials and Methods

Observers and experimental setup

Thirteen human observers (18–35 years of age, five males and eight females recruited from students and staff at New York University) participated in the experiments. Observers had normal or corrected-to-normal vision. They were naive to the purpose of the experiment, except for one observer who is an author (G.O.). They all provided informed written consent before participation. All experimental procedures were approved by the Institutional Review Board at New York University.

Throughout the experiment, subjects were seated in an adjustable chair in a semidark room with chin and forehead supported before a CRT display monitor (21-inch Sony GDM-5402; 75 Hz refresh rate; 1600 × 1200 pixels screen resolution; 52 cm viewing distance). Stimulus presentation was controlled with the Psychophysics Toolbox (Brainard, 1997) and MATLAB (MathWorks). Eye movements were monitored using a high-speed infrared camera (EyeLink, SR Research). Gaze position was recorded at 1 kHz.

Experimental design

Stochastic multifeature face categorization task. The task required the classification of faces into two categories, each defined by a prototype face (Fig. 1*A*,*B*). The subject initiated each trial by fixating a small red point at the center of the screen [fixation point (FP), 0.3° diameter]. After a short delay (200–500 ms, truncated exponential distribution), two targets appeared 5° above and below the FP to indicate the two possible face category choices (category 1 or 2). Simultaneously with the target onset, a face stimulus (2.18° × 2.83°, ~83 × 108 pixels) appeared on the screen parafoveally (stimulus center 1.8° to the left or right of the FP, counterbalanced across subjects; results were similar for the two sides). We placed the stimuli parafoveally, aiming to present the informative facial features at comparable visual eccentricities and yet keep the stimuli close enough to the fovea to take advantage of the foveal bias for face perception (Levy et al., 2001; Kreichman et al., 2020). The parafoveal presentation also enabled us to control subjects' fixation so that small eye movements (e.g., microsaccades) within the acceptable fixation window did not substantially change the sensory inputs. Subjects reported the face category by making a saccade to one of the two targets as soon as they were ready. The stimulus was extinguished immediately after the saccade initiation. Reaction times were calculated as the time from the stimulus onset to the saccade initiation. If subjects failed to make a choice in 5 s, the trial was aborted (0.101% of trials). To manipulate task difficulty, we created a morph continuum between the two prototypes and presented intermediate morphed faces on different trials (see below). Distinct auditory feedbacks were delivered for correct and error choices. When the face was ambiguous (halfway between the two prototypes on the morph continuum), correct feedback was delivered on a random half of trials.

Subjects could perform two categorization tasks, identity categorization (Fig. 1*B*, top) and expression categorization (Fig. 1*B*, bottom). The prototype faces for each task were chosen from the photographs of MacBrain Face Stimulus Set (Tottenham et al., 2009). For the illustrations of identity stimuli in Figure 1, *A*, *B*, and *C*, we used morphed images of two authors' faces to avoid copyright issues. We developed a custom algorithm that morphed different facial features (regions of the stimulus) independently between the two prototype faces. Our algorithm started with 97–103 manually matched anchor points on the prototypes and morphed one face into another by linear interpolation of the positions of anchor points and textures inside the tessellated triangles defined by the anchor points. The result was a perceptually seamless transformation of the geometry and internal features from one face to another. Our method enabled us to morph different regions of the faces independently. We focused on three key regions (eyes, nose, and mouth) and created an independent series of morphs for each one of them. The faces that were used in the task were composed of different morph levels of these three informative features. Anything outside those features was set to the halfway morph between the prototypes and thus was uninformative. The informativeness of the three features (stimulus strength) was defined based on the mixture of prototypes, spanning from −100% when the feature was identical to prototype 1 to +100% when it was

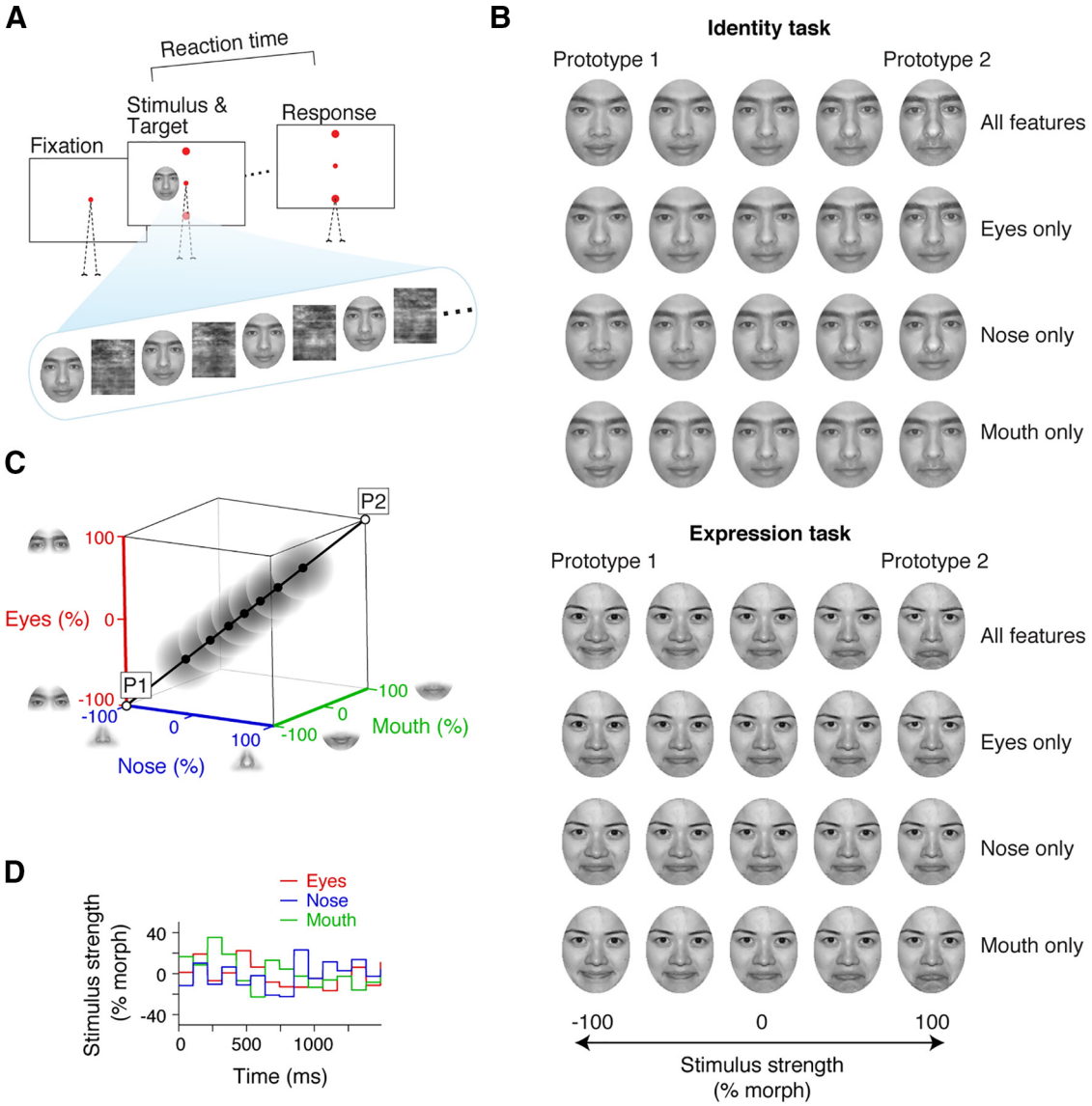
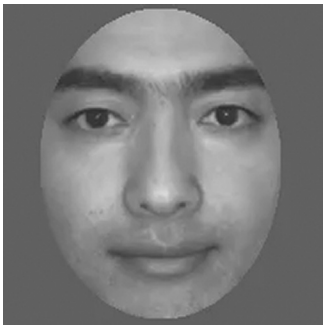


Figure 1. Stochastic multifeature face categorization task. ***A***, On each trial, after the subject fixated on a central fixation point, two circular targets appeared on the screen and were immediately followed by a dynamic stimulus stream consisting of faces interleaved with masks (Movie 1). The subject was instructed to report the stimulus category (face identity or expression in different blocks) as soon as ready by making a saccadic eye movement to one of the two targets associated with the two categories. Reaction time was defined as the interval between the stimulus onset and the saccade onset. ***B***, The stimuli in each task were engineered by morphing two category prototype stimuli, defined as ±100% stimulus strengths. The prototype faces were chosen from the photographs of MacBrain Face Stimulus Set (Tottenham et al., 2009). Our custom algorithm allowed independent morphing of different stimulus regions. We defined three regions (eyes, nose, and mouth) as informative features while fixing other regions to the intermediate level between the prototypes. The top row for each task shows a morph continuum where all facial features vary together between the two prototypes. The bottom three rows show the morph continua for individual informative features while keeping the other features at the intermediate level between the prototypes (0% morph level). For the identity stimuli shown in the figure, we used faces of two authors (G.O. and R.K.) to avoid copyright issues. ***C***, Because we independently morphed the three facial features, the stimulus space for each task was three dimensional. In this space, the prototypes (P1 and P2) are two opposite corners. In each trial, the nominal mean stimulus strength was sampled from points on the diagonal line (filled black dots). Each stimulus frame was then sampled from a 3D symmetric Gaussian distribution around the specified nominal mean (gray clouds; SD 20% morph). ***D***, During the stimulus presentation in each trial, the morph levels of the three informative features were randomly and independently sampled from the Gaussian distribution every 106.7 ms. The features therefore furnished differential and time-varying degrees of support for the competing choices in each trial. Each face in the stimulus stream was masked with a random noise pattern created by phase scrambling images of faces (***A***). The masking prevented subjects from detecting the feature changes, although these changes still influenced the subjects' final choices. The stimuli were presented parafoveally to keep the informative features at more or less similar eccentricities in the visual field.

identical to prototype 2 (Fig. 1*C*). At the middle of the morph line (0% morph), the feature was equally shaped by the two prototypes.

By varying the three features independently, we could study spatial integration through creating ambiguous stimuli in which different features could support different choices (Fig. 1*C*). We could also study temporal integration of features by varying the three discriminating features every 106.7 ms within each trial (Fig. 1*D*). This frame duration provide us with sufficiently precise measurements of subjects' temporal integration in their ~1 s decision times while ensuring the smooth subliminal transition of frames (see below). The stimulus strengths of three features in each trial were drawn randomly from independent Gaussian distributions. The mean and SD of these distributions were equal and fixed within each trial, but the means varied randomly from trial to trial. For the identity task, we tested the following seven mean stimulus strengths: –50%, –30%, –14%, 0%, +14%, +30%, and +50%. For the expression task, we used –50%, –20%, –10%, 0%, +10%, +20%, and +50%, except for subject 13, who had a higher behavioral threshold and was also exposed to ±80% morph levels. The SD was 20%



Movie 1. The video shows an example image sequence similar to those used in the experiments. The sequence consists of face images interleaved by masks. For each face image, the morph levels of three facial features (eyes, nose, mouth) were randomly sampled from a Gaussian distribution centered on the nominal morph level for the trial (Fig. 1*C*). The masks made these stimulus fluctuations subliminal. Note that the size and frame rate of the video does not reproduce the actual stimuli used in the experiments. [View online]

morph level. Sampled values that fell outside the range of −100% to +100% (0.18% of samples) were replaced with new samples inside the range. Using larger SDs would have allowed us to sample a wider stimulus space, but we limited the SD to 20% morph level to keep the stimulus fluctuations subliminal, avoiding potential changes of decision strategy for vividly varying stimuli.

Changes in the stimulus within a trial were implemented in a subliminal fashion so that subjects did not consciously perceive variation of facial features, and yet their choices were influenced by these variations. We achieved this goal using a sequence of stimuli and masks within each trial (Movie 1). The stimuli were morphed faces with a particular combination of the three discriminating features. The masks were created by phase randomization (Heekeren et al., 2004) of the 0% morph face and therefore had largely matching spatial frequency content with the stimuli shown in the trial. The masks ensured that subjects did not consciously perceive minor changes in informative features over time within a trial. In debriefings following each experiment, subjects noted that they saw one face in each trial, but the face was covered with time-varying cloudy patterns (i.e., masks) over time.

For the majority of subjects (9 of 13), each stimulus was shown without a mask for one monitor frame (13.3 ms). Then, it gradually faded out over the next seven frames as a mask stimulus faded in. For these frames, the mask and the stimulus were linearly combined, pixel by pixel, according to a half-cosine weighting function, so that in the last frame, the weight of the mask was 1 and the weight of the stimulus was 0. Immediately afterward, a new stimulus frame with a new combination of informative features was shown, followed by another cycle of masking, and so on. For a minority of subjects (4 of 13), we replaced the half-cosine function for the transition of stimulus and mask with a full-cosine function, where each eight-frame cycle started with a mask, transitioned to an unmasked stimulus in frame 5, and transitioned back to a full mask by the beginning of the next cycle. We did not observe any noticeable difference in the results of the two presentation methods and combined data across subjects.

Twelve subjects participated in the identity categorization task (35,300 total trials; mean ± SD trials per subject, 2942 ± 252). Seven subjects participated in the expression categorization task in separate sessions (20,225 total trials; trials per subject, 2889 ± 285). Six of the subjects performed both tasks. Our subject counts are comparable to previous studies of perceptual decision-making tasks (Levi et al., 2018; Stine et al., 2020). Collecting a large number of trials from individual subjects enabled detailed quantification of decision behavior for each subject (Smith and Little, 2018). Our results were highly consistent across subjects. A part of the data for the identity categorization task was previously published (Okazawa et al., 2018).

Odd-one-out discrimination task. Our behavioral analyses and decision-making models establish that subjects' choices in the identity and expression categorization tasks were differentially informed by the three facial features; choices were most sensitive to changes in the morph level of eyes for identity discrimination and changes in the morph level of mouth for expression discrimination (Fig. 2*E*,*F*). This task-dependent sensitivity to features could arise from two sources: different visual discriminability for the same features in the two tasks and/or unequal decision weights for informative features in the two tasks (see Fig. 10*A*). To determine the relative contributions of these factors, we designed an odd-one-out discrimination task to measure visual discriminability of different morph levels of informative features in the two tasks (see Fig. 10*B*).

On each trial, subjects viewed three stimuli presented sequentially at 1.8° eccentricity (similar to the categorization tasks). The stimuli appeared after the subject fixated a central FP, shown for 320 ms each, with 500 ms interstimulus intervals. The three stimuli in a trial were the same facial feature (eyes, nose, or mouth) but had distinct morph levels, chosen randomly from the following set: −100%, −66%, −34%, 0%, +34%, +66%, +100%. Facial regions outside the target feature were masked by the background. The target feature varied randomly across trials. Subjects were instructed to report the odd stimulus in the sequence (the stimulus most distinct from the other two) by pressing one of the three response buttons within 2 s from the offset of the last stimulus (RT from stimulus offset, $0.66 \pm 0.13\,s$, mean ± SD). No feedback was given after the response. Subjects underwent extensive training before the data collection to achieve stable and high performance. During training, two of the three stimuli were identical, and subjects received feedback on whether they correctly chose the distinct stimulus. The training continued until subjects reached 70% correct choices (chance level 33%).

Nine of the 12 subjects who participated in the identity categorization task also performed the odd-one-out discrimination task using identity stimuli in separate blocks of the same sessions. Three of the seven subjects who participated in the expression task performed the odd-one-out task using expression stimuli. For the identity stimuli, 13,648 trials were collected across the three features (nine subjects, 1516 ± 420 trials per subject, mean ± SD). For the expression stimuli, 3570 trials were collected (three subjects, 1190 ± 121 trials per subject).

Single-feature categorization task. As an alternative method to quantify the visual discriminability for individual facial features, we also performed a single-feature categorization task with a subset of subjects (see Fig. 11*A*). In this task, the subjects categorized the facial identities as in the main identity categorization task but based their decisions on only one facial feature shown on each trial. Facial regions outside the target feature were replaced by the background. The task structure was the same as that of the main task. Trials of the three facial features were randomly interleaved. To capture the full extent of psychometric functions, we used morph levels ranging from −150% to +150% (see Fig. 11*B*). The stimuli beyond 100% indicate extrapolation from the prototypes, but the extrapolated images looked natural within the tested range.

Four of the 12 subjects who performed the main identity categorization task also performed the single-feature task in the same sessions. We collected in total 5571 trials (1393 ± 117 trials per subject, mean ± SD).

Statistical analysis

Psychometric and chronometric functions. We assessed the effects of stimulus strength on the subject's performance by using logistic regression (Fig. 2*A*,*B*) as follows:

$$logit[P(choice2)] = \alpha_0 + \alpha_1 s, \quad (1)$$

where $logit(p) = log(p/1-p)$, s is the nominal stimulus strength ranging from −1 (−100% morph level) to +1 (+100% morph level), and α_i are regression coefficients; α_0 quantifies the choice bias and α_1 quantifies the slope of the psychometric function.

The relationship between the stimulus strength and the subject's mean RTs was assessed using a hyperbolic tangent function (Shadlen et al., 2006; Fig. 2*C*,*D*) as follows:

$$T = \frac{\beta_0}{s} tanh(\beta_1 s) + \beta_2, \quad (2)$$

where T is the mean RTs measured in milliseconds and β_i are model parameters; β_0 and β_1 determine the stimulus-dependent changes in

decision time, whereas β_2 quantifies the sensory and motor delays that elongate the RTs but are independent of the decision-making process (i.e., nondecision time).

Psychophysical Reverse Correlation. To quantify the effect of stimulus fluctuations over time and space (facial features) on choice (Fig. 1*D*), we performed psychophysical reverse correlation (Ahumada, 1996; Okazawa et al., 2018). Psychophysical kernels $K_f(t)$ were calculated as the difference of average fluctuations of morph levels conditional on the subject's choices as follows:

$$K_f(t) = E[s_f(t)|choice\,1] - E[s_f(t)|choice\,2], \quad (3)$$

where $s_f(t)$ is the morph level of feature f at time t. This analysis only used trials with low stimulus strength (nominal morph level, 0–14%; 14,213 trials across 12 subjects in the identity task and 7882 trials across seven subjects in the expression task). For the nonzero strength trials, the mean strength was subtracted from the fluctuations, and the residuals were used for the reverse correlation. For the time course of psychophysical kernels (see Figures 6; 8, *D* and *E*; and 9, *E* and *F*), we use stimulus fluctuations up to the median RT aligned to stimulus onset or the saccade onset to ensure that at least half the trials contributed to the kernels at each time. Figure 2, *E* and *F*, shows the kernels averaged over time from stimulus onset to median RT. We did not perform any smoothing on the kernels of aggregated data (see Figs. 2, *E* and *F*; 6, *A* and *B*; 8, *D* and *E*; and 9, *E* and *F*). For individual subjects' kernels (see Fig. 6*C*,*D*), we applied three-point boxcar smoothing to reduce noise.

Joint psychometric function. To quantify the effect that cofluctuations of feature strengths have on choice, we quantified the probability of choices as a function of the joint distribution of the stimulus strengths across trials (Fig. 3*A*,*B*). We constructed the joint distribution of the three features by calculating the average strength of each feature in the trial. Thus, one trial corresponds to a point in a 3D feature space (Fig. 1*C*). In this space, the probability of choice was computed within a Gaussian window with a SD of 4%. Figure 3, *A* and *B*, shows 2D intersections of this 3D space. We visualized the probability of choice by drawing iso-probability contours at 0.1 intervals. The trials of all stimulus strengths were included in this analysis, but similar results were also obtained by restricting the analysis to the low morph levels (≤14%). We aggregated data across all subjects, but similar results were observed within individual subjects.

To quantify linear and multiplicative effects on joint psychometric functions, we performed the following logistic regression:

$$\begin{aligned} logit[P(choice2)] = \; & w_e s_e + w_n s_n + w_m s_m + \\ & w_{e,n} s_e s_n + w_{n,m} s_n s_m + w_{m,e} s_m s_e + , \\ & w_{e,n,m} s_e s_n s_m \end{aligned} \quad (4)$$

where s_e, s_n, and s_m corresponds to the stimulus strengths of eyes, nose, and mouth averaged within individual trials; w_e, w_n, and w_m are model coefficients for linear factors, whereas $w_{e,n}$, $w_{n,m}$, $w_{m,e}$, and $w_{e,n,m}$ are coefficients for multiplicative factors. In this regression, the dynamic ranges of the linear and multiplicative terms were scaled to match to ensure a more homogeneous distribution of explainable variance across different factors; otherwise, the fluctuations of multiplicative terms would be one or two orders of magnitude smaller than those of the linear terms. Figure 3, *C* and *D*, shows the coefficients averaged across subjects.

Relationship between stimulus strength and subjective evidence. To quantitatively predict behavioral responses from stimulus parameters, one must first know the mapping function between the physical stimulus strength (morph level) and the amount of evidence subjects acquired from the stimulus. This mapping could be assessed by performing a logistic regression that relates choice to different ranges of stimulus strength (Fig. 4), similar to those performed in previous studies (Yang and Shadlen, 2007; Waskom and Kiani, 2018). For this analysis, we used the following regression:

$$logit[P(choice2)] = \sum_f \sum_{k \in bins} w_{f,k} N_{f,k}, \quad (5)$$

where $N_{f,k}$ is the number of stimulus frames that fall into decile k for feature f in a given trial, and $w_{f,k}$ are regression coefficients that signify the subjective evidence assigned to a morph level k of feature f. Division of feature morph levels into deciles in our regression aimed at limiting the number of free parameters while maintaining adequate resolution to quantify the mapping function between the stimulus strength and momentary evidence. If the subjective evidence in units of log-odds scales linearly with the morph level, then $w_{f,k}$ would linearly change with the morph level. Plotting $w_{f,k}$ as a function of feature morph level indicated a linear relationship, except perhaps for the extreme deciles at the two ends of the morph line (Fig. 4). For illustration purposes, the fitting lines in Figure 4 exclude the extreme deciles, but to ensure unbiased reporting of statistics, we included all the deciles to quantify the accuracy of the linear fit (see below, Results).

Model fit and evaluation

To quantitatively examine the properties of the decision-making process, we fit several competing models to the subject's choices and RTs. Based on our earlier analyses (Figs. 3, 4), these models commonly use a linear mapping between feature morph levels and the evidence acquired from each feature, as well as linear functions for spatial integration of informative features in each frame. The combined momentary evidence from each stimulus frame was then integrated over time. Our main models are therefore extensions of the drift diffusion model, where fluctuations of the three informative facial features are accumulated toward decision bounds, and reaching a bound triggers a response after a nondecision time. Our simplest model used linear integration over time, whereas our more complex alternatives allowed leaky integration or dynamic changes of sensitivity over time. We also examined models that independently accumulate the evidence from each informative feature (i.e., three competing drift diffusion processes), where the decision and RT were determined by the first process reaching a bound. Below, we first provide the equations and intuitions for the simplest model and explain our fitting and evaluation procedures. Afterward, we explain the alternative models.

Spatial integration in our models linearly combines the strength of features at each time to calculate the momentary evidence conferred by a stimulus frame as follows:

$$\mu(t) = k_e s_e(t) + k_n s_n(t) + k_m s_m(t), \quad (6)$$

where $s_e(t)$, $s_n(t)$, $s_m(t)$ are the strengths of eyes, nose, and mouth at time t, and k_e, k_n, k_m are the sensitivity parameters for each feature. Momentary evidence ($\mu(t)$) was integrated over time to derive the decision variable. The process stopped when the decision variable reached a positive or negative bound ($\pm B$). The probability of crossing the upper and lower bounds at each decision time can be calculated by solving the Fokker-Planck equation (Karlin and Taylor, 1981; Kiani and Shadlen, 2009) in the following:

$$\frac{\delta p(v,t)}{\delta t} = \left[-\frac{\delta}{\delta v}\mu(t) + 0.5\frac{\delta^2}{\delta v^2}\sigma^2 \right] p(v,t), \quad (7)$$

where $p(v,t)$ is the probability density of the decision variable at different times. The boundary conditions are as follows:

$$\begin{aligned} p(v,0) &= \delta(v) \\ p(\pm B, t) &= 0 \end{aligned}, \quad (8)$$

where $\delta(v)$ denotes a delta function. The first condition enforces that the decision variable always starts at zero, and the second condition guarantees that the accumulation terminates when the decision variable reaches one of the bounds. We set the diffusion noise (σ) to 1 and defined the bound height and drift rate in units of σ. RT distribution for each choice was obtained by convolving the distribution of bound crossing times with the distribution of nondecision time, which was defined as a Gaussian distribution with a mean of T_0 and an SD of σ_{T_0}.

Overall, this linear integration model had six degrees of freedom: decision bound height (B), sensitivity parameters (k_e, k_n, k_m), and the

parameters for nondecision time (T_0, σ_{T_0}). We fit model parameters by maximizing the likelihood of the joint distribution of the observed choices and RTs in the experiment (Okazawa et al., 2018). For a set of parameters, the model predicted the distribution of RTs for each possible choice for the stimulus strengths used in each trial. These distributions were used to calculate the log likelihood of the observed choice and RT on individual trials. These log likelihoods were summed across trials to calculate the likelihood function for the dataset. Model parameters were optimized to maximize this function. To avoid local maxima, we repeated the fits from 10 random initial points and chose the fit with the highest likelihood. The trials of all stimulus strengths were included in this fitting procedure. The fits were performed separately for each subject. Figure 5, *B* and *C*, shows the average fits across subjects.

To generate the model psychophysical kernels, we created 10^5 test trials with 0% stimulus strength using the same stimulus distributions as in the main task (i.e., Gaussian distribution with a SD of 20%). We simulated model responses for these trials with the same parameters fitted for each subject. We then used the simulated choices and RTs to calculate the model prediction for psychophysical kernels of the three features. Figure 6, *A* and *B*, shows the average predictions superimposed on the observed psychophysical kernels, averaged across subjects (Fig. 6*C*,*D*, single-subject example). Note that the model kernels were not directly fit to match the data. They were calculated based on an independent set of simulated 0% stimulus trials, making the comparison in Figure 6 informative.

The same fitting procedure was used for the alternative models explained below.

Leaky integration. To test the degree of temporal integration, we added a memory loss (leak) in the decision-making process. This model is implemented as an Ornstein-Uhlenbeck process, whose Fokker-Planck equation is the following:

$$\frac{\delta p(v,t)}{\delta t} = \left[\frac{\delta}{\delta v}\left(\lambda v - \mu(t)\right) + 0.5\frac{\delta^2}{\delta v^2}\sigma^2\right]p(v,t), \tag{9}$$

where λ is the leak rate. A larger leak rate indicates greater loss of information over time. At the limit of an infinitely large leak rate, the model no longer integrates evidence and makes a decision based solely on whether the most recently acquired momentary evidence exceeds one of the decision bounds.

Dynamic sensitivity. To test whether the effect of sensory evidence on choice is constant over time, we allowed sensitivity to features to be modulated dynamically. To capture both linear and nonlinear temporal changes, the modulation included linear (γ_1) and quadratic (γ_2) terms as follows:

$$\mu(t) = \left(k_e s_e(t) + k_n s_n(t) + k_m s_m(t)\right) \times (1 + \gamma_1 t + \gamma_2 t^2). \tag{10}$$

Parallel accumulation of evidence from three facial features. The models above first integrated the evidence conferred by the three informative facial features (spatial integration) and then accumulated this aggregated momentary evidence over time. We also considered alternative models, in which evidence from each feature was accumulated independently over time. These models therefore included three competing accumulators. Each accumulator received momentary evidence from one feature with fixed sensitivity (k_e, k_n, k_m for eyes, nose, mouth). In the model (see Fig. 8*A*), the accumulator that first reached a decision bound dictated the choice and decision time. As in the models above, a nondecision time separated the bound crossing and response time.

To further explore different decision rules, we constructed two variants of the parallel accumulation model (see Fig. 9*A*,*B*). In the first variant, the decision was based on the sign of the majority of the accumulators (i.e., two or more of three accumulators) at the moment when one accumulator reached the bound. In the second variant, the decision was based on the sign of the sum of the decision variables across the three accumulators at the time when one accumulator reached the bound. All model variants had six free parameters ($k_e, k_n, k_m, B, T_0, \sigma_{T_0}$), equal to the degrees of freedom of the main model explained above.

Analysis of odd-one-out discrimination task

We used subjects' choices in the odd-one-out task to estimate visual discriminability of different morph levels of the informative features. We adopted an ideal observer model developed by Maloney and Yang (2003), where the perception of morph level i of feature f is defined as a Gaussian distribution, $\mathcal{N}(\psi_{i,f}, \sigma_f)$, with mean $\psi_{i,f}$ and SD σ_f. The discriminability of a triad of stimuli (i, j, k) is determined by perceptual distances $(|\psi_{i,f} - \psi_{j,f}|, |\psi_{j,f} - \psi_{k,f}|, |\psi_{k,f} - \psi_{i,f}|)$. Specifically, an ideal observer performing the odd-one-out discrimination of three morph levels (i, j, k) of feature f would choose i if the perceptual distances of i from j and k are larger than the perceptual distance of j and k as in the following:

$$p(choice = i) = P(|\psi_{i,f} - \psi_{j,f}| - |\psi_{j,f} - \psi_{k,f}| > 0) \cdot P(|\psi_{i,f} - \psi_{k,f}| - |\psi_{j,f} - \psi_{k,f}| > 0). \tag{11}$$

The probabilities on the right side of the equation can be derived from $\mathcal{N}(\psi_{i,f}, \sigma_f)$. The probability of choosing j and k can be calculated in a similar way (Maloney and Yang, 2003). The model captures both the discriminability of morph levels of the same feature through $\psi_{i,f}$, and differences across features through σ_f.

We fit the ideal observer model to the subject's choices using maximum likelihood estimation. As there were seven morph levels for each feature in our task, choices for each feature could be explained using eight parameters ($\psi_{i,f}$ for $i = 1 \ldots 7$, and σ_f). Six of these parameters are free ($\psi_{2,f} \ldots \psi_{6,f}$ and σ_f), and $\psi_{1,f}$ and $\psi_{7,f}$ were anchored at −1 and +1, respectively, to avoid redundant degrees of freedom in the fits. The model was fit separately for each subject and feature. To avoid local maxima, we repeated the fits from 10 random initial points and chose the parameters that maximized the likelihood function. Because $\psi_{i,f}$ changed largely linearly with the feature strength (% morph) in all fits (see Fig. 10*D*) and the range of $\psi_{i,f}$ was fixed at [−1, +1], we could quantify perceptual discriminability of different features the respective σ_f. Specifically, d' between the two extreme morph levels of a feature is $2/\sigma_f$ as follows:

$$d'_f = (\psi_{7,f} - \psi_{1,f})/\sigma_f = 2/\sigma_f. \tag{12}$$

If the differences of feature sensitivity parameters (k_e, k_n, k_m; Eq. 6) in the categorization tasks were fully determined by the visual discriminability of features, that is, there was no task-dependent weighting of the features, k_f would have the same relative scales as the $1/\sigma_f$. To test this, we divided the model sensitivities for the three facial features by $1/\sigma_f$ for each to estimate task-dependent decision weights. The resulting decision weights showed significant inhomogeneity across features and between tasks (Fig. 10*G*), suggesting the presence of task-dependent weighting of facial features during categorization.

Results

Spatial integration in face categorization

We developed stochastic multifeature face categorization tasks suitable for studying spatial and temporal properties of the computations underlying the decision-making process. Subjects classified naturalistic face stimuli into two categories. In each trial, subjects observed a face stimulus with subliminally varying features and, when ready, reported the category with a saccadic eye movement to one of the two targets (Fig. 1*A*). The targets were associated with the two prototypes that represented the discriminated categories—identities of two different people in the identity categorization task (Fig. 1*B*, top) or happy and sad expressions of a person in the expression categorization task (Fig. 1*B*, bottom). The stimulus changed dynamically in each

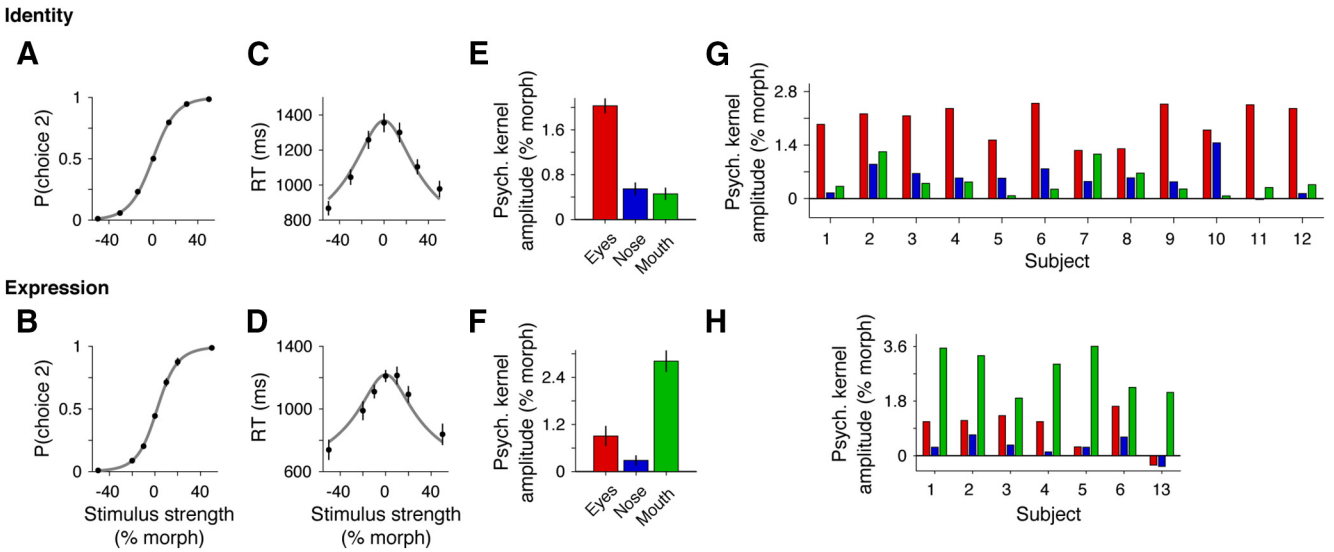


Figure 2. Stimulus strength shaped choices and reaction times with differential contributions from informative features. ***A***, ***B***, Psychometric functions based on nominal stimulus strength in the identity (***A***) and expression (***B***) categorization tasks. Gray lines are logistic fits (Eq. 1). Error bars are SEM across subjects (most of them smaller than data dots). ***C***, ***D***, Chronometric functions based on nominal stimulus strength in the two tasks. Average reaction times were slower for the intermediate, most ambiguous stimulus strengths. Gray lines are the fits of a hyperbolic tangent function (Eq. 2). ***E***, ***F***, Psychophysical reverse correlation using feature fluctuations revealed positive but nonuniform contributions of multiple features in both tasks. The amplitude of the psychophysical kernel for a feature was calculated as the difference of the average feature fluctuations conditioned on the two choices. Trials with 0–14% nominal stimulus strength were used in this analysis. Error bars indicate SEM across subjects. ***G***, ***H***, The kernel amplitudes for individual subjects. Most subjects showed positive kernels for multiple facial features.

trial. The dynamic stimulus stream consisted of a sequence of face stimuli interleaved by masks (Fig. 1*A*). Each face stimulus was engineered to have three informative features in the eyes, nose, and mouth regions, and sensory evidence conferred by the three informative features rapidly fluctuated over time as explained in the next paragraph. The masks between face stimuli kept the changes in facial features subliminal, creating the impression that a fixed face was covered periodically with varying noise patterns (see Materials and Methods; Movie 1).

Using a custom algorithm, we could independently morph the informative facial features (eyes, nose, mouth) between the two prototypes (Fig. 1*B*) to create a 3D stimulus space whose axes correspond to the morph level of the informative features (Fig. 1*C*). In this space, the prototypes are two opposite corners (specified as ± 100% morph), and the diagonal connecting the prototypes correspond to a continuum of faces where the three features of each face equally support one category versus the other. For the off-diagonal faces, however, the three features provide unequal or even opposing information for the subject's choice. In each trial, the nominal mean stimulus strength (% morph) was sampled from the diagonal line (Fig. 1*C*, black dots). The dynamic stimulus stream was created by independently sampling a stimulus every 106.7 ms from a 3D symmetric Gaussian distribution with the specified nominal mean (SD 20% morph; Fig. 1*C*,*D*). The presented stimuli were therefore frequently off diagonal in the stimulus space. The subtle fluctuations of features influenced subjects' choices as we show below, enabling us to determine how subjects combined spatiotemporal sensory evidence over space and time for face categorization.

We first evaluated subjects' choices and RTs in both tasks (Fig. 2). The average correct rate excluding 0% morph level was 91.0% ± 0.7% (mean ± SEM across subjects) for the identity task and 89.2% ± 1.2% for the expression task. The choice accuracy monotonically improved as a function of the nominal mean stimulus strength in the trial (Fig. 2*A*,*B*; identity task, $\alpha_1 = 9.6 \pm 1.8$ in Eq. 1; $t_{(11)} = 18.3$, $p = 1.4 \times 10^{-9}$; expression task, $\alpha_1 = 11.3 \pm 3.0$, $t_{(6)} = 9.8$, $p = 6.4 \times 10^{-5}$, two-tailed t test). Correspondingly, the reaction times became faster for higher stimulus strengths (Fig. 2*C*,*D*; identity task, $\beta_1 = 4.7 \pm 1.0$ in Eq. 2; $t_{(11)} = 16.4$, $p = 4.4 \times 10^{-9}$; expression task, $\beta_1 = 4.9 \pm 1.9$, $t_{(6)} = 6.6$, $p = 5.6 \times 10^{-4}$). These patterns are consistent with evidence accumulation mechanisms that govern perceptual decisions with simpler stimuli, for example, direction discrimination of random dots motions (Smith and Vickers, 1988; Ratcliff and Rouder, 2000; Palmer et al., 2005). However, the decision-making mechanisms that do not integrate sensory evidence over time can also generate qualitatively similar response patterns (Waskom and Kiani, 2018; Stine et al., 2020). Furthermore, because in our task design we used identical nominal mean morph levels for the informative features in a trial, characterizing behavior based on the mean levels cannot reveal if subjects integrated sensory evidence across facial features (spatial integration). However, we can leverage the stochastic fluctuations of the stimulus to test whether sensory evidence was integrated over space and time. In what follows, we first quantify the properties of spatial integration and then examine the properties of temporal integration.

To test whether multiple facial features informed subjects' decisions, we used psychophysical reverse correlation to evaluate the effect of the fluctuations of individual features on choice. Psychophysical kernels were generated by calculating the difference between the feature fluctuations conditioned on choice (Eq. 3). We focused on trials with the lowest stimulus strengths where choices were most strongly influenced by the feature fluctuations (0–14%; mean morph level of each trial was subtracted from the fluctuations; see Materials and Methods). Figure 2, *E* and *F*, shows the kernel amplitude of the three facial features averaged over time from stimulus onset to median RT. These kernel amplitudes quantify the overall sensitivity of subjects' choices to equal fluctuations of the three features (in %morph units). The kernel amplitudes markedly differed across features in each task (identity task: $F_{(2,33)} = 55.4$, $p = 2.8 \times 10^{-11}$,

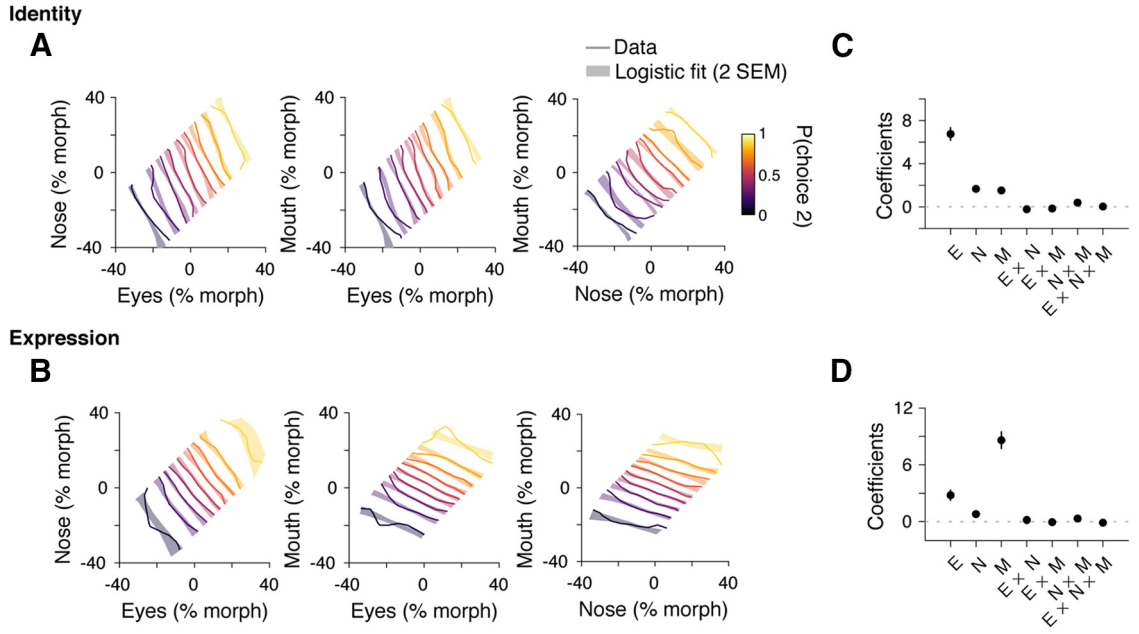


Figure 3. Spatial integration across informative features was largely linear. ***A***, ***B***, Two-dimensional slices of the psychometric function for the three-dimensional stimulus space. The iso-probability contours are shown for the probability of choice 2 as a function of the true average morph level of two of the three informative features in each trial. The third informative feature was marginalized for ***A*** and ***B***. The iso-probability contours are drawn at 0.1 intervals and moderately smoothed with a 2D Gaussian function (SD 4% morph). Thin lines are actual data, and thick pale lines are a logistic fit (Eq. 4; the thickness of the lines reflects 2 SEM of the fitted parameters). The straight and parallel contour lines are compatible with a largely linear spatial integration process across features, with the slope of the contours reflecting the relative contribution of the features illustrated in each panel. ***C***, ***D***, A logistic regression to evaluate the relative contribution of individual features and the multiplicative interactions to subjects' choices. The regression coefficients supported a largely linear spatial integration process; the interaction terms have minimal impact on choice beyond the linear integration across features. E, eyes; N, nose; M, mouth. Error bars indicate SEM across subjects.

expression task: $F_{(2,18)} = 33.6$, $p = 8.5 \times 10^{-7}$; one-way ANOVA), greatest for the eyes region in the identity task ($p < 9.5 \times 10^{-10}$ compared with nose and mouth, *post hoc* Bonferroni test) and for the mouth region in the expression task ($p < 3.9 \times 10^{-5}$ compared with eyes and nose).

Critically, the choice was influenced by more than one feature. In the identity task, all three features had significantly positive kernel amplitudes (eyes, $t_{(11)} = 15.4$, $p = 8.6 \times 10^{-9}$; nose, $t_{(11)} = 4.8$, $p = 5.4 \times 10^{-4}$; mouth, $t_{(11)} = 4.2$, $p = 0.0015$, two-tailed t test for each feature). In the expression task, mouth and eyes had statistically significant kernel amplitudes, and nose had a positive kernel, although it did not reach significance (mouth, $t_{(6)} = 10.3$, $p = 4.8 \times 10^{-5}$; eyes, $t_{(6)} = 3.6$, $p = 0.012$; nose, $t_{(6)} = 2.2$, $p = 0.072$). Positive kernels for multiple facial features were prevalently observed for individual subjects too (Fig. 2*G*,*H*). Therefore, the pooled results are not because of mixing data from multiple subjects with distinct behavior. These results suggest that subjects use multiple facial features for categorization, but the features nonuniformly contribute to their decisions, and their relative contributions differ between the tasks (interaction between feature kernels and tasks: $F_{(2,10)} = 90.5$, $p = 3.9 \times 10^{-7}$; two-way repeated measures ANOVA with six subjects who performed both tasks), which we revisit in the following sections.

Although the amplitude of psychophysical kernels informs us about the overall sensitivity of choice to feature fluctuations in the face stimuli, it does not clarify the contribution of sensory and decision-making processes to this sensitivity. Specifically, subjects' choices may be more sensitive to changes in one feature because the visual system is better at discriminating the feature changes (visual discriminability) or because the decision-making process attributes a larger weight to the changes of that feature (decision weights; Schyns et al., 2002; Sigala and Logothetis, 2002). We dissociate these factors in the final section of Results, but for the intervening sections, we focus on the overall sensitivity of choice to different features.

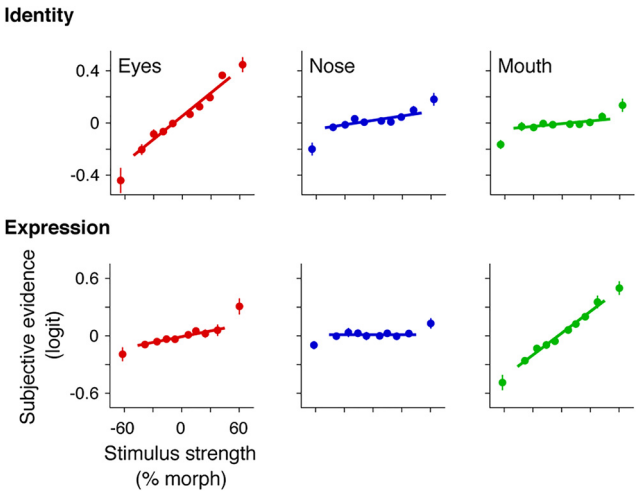


Figure 4. The evidence conferred by a feature mapped linearly to the feature morph level, especially for the intermediate stimuli. The linear integration of features across space allowed accurate quantification of the evidence that subjects inferred from each feature. We split morph levels of each feature into 10 quantiles and used a logistic regression that explained choices based on the number of occurrences of the quantiles in each trial, quantifying the subjective evidence of each morph level in units of log odds of choice. Subjective evidence linearly scaled with the morph level for each feature, except for the highest strengths. The lines are linear regressions of subjective evidence against feature morphs, excluding the highest strengths. Error bars indicate SEM across subjects.

Linearity of spatial integration of facial features

How do subjects integrate information from multiple spatial features? Could it be approximated as a linear process, or does it involve significant nonlinear effects, for example, synergistic interactions that magnify the effect of cofluctuations across features? Nonlinear effects can be empirically discovered by plotting joint psychometric functions that depict subjects' accuracy as a function of the strength of the facial features (Fig. 3*A*,*B*). Here, we define the true mean strength of each feature as the average of the feature morph levels over the stimulus frames shown on each trial (see Figs. 6, 7 for temporal effects). The plots visualize the three orthogonal 2D slices of the 3D stimulus space (Fig. 1*C*), and the contour lines show the probability of choosing the second target (choice 2) at the end of the trial.

These iso-performance contours (Fig. 3*A*,*B*, thin lines) were largely straight and parallel to each other, suggesting that a weighted linear integration across features underlies behavioral responses. The slope of contours in each 2D plot reflects the relative contribution of the two facial features to choice. For example, the nearly vertical contours in the eyes-versus-nose plot of the identity task indicate that eyes had a much greater influence on subjects' choices, consistent with the amplitudes of psychophysical kernels (Fig. 2*E*). Critically, the straight and parallel contour lines indicate that spatial integration does not involve substantial nonlinearity. A linear model, however, does not explain curved contours, which appear at the highest morph levels, especially in the 2D plots of the less informative pairs (e.g., the nose × mouth plot for the identity task). Multiple factors could give rise to the curved contours. First, subjects rarely make mistakes at the highest morph levels, reducing our ability to perfectly define the contour lines at those levels. Second, the 2D plots marginalize over the third informative feature, and this marginalization is imperfect because of finite trial counts in the dataset. Put together, we cannot readily attribute the presence of curved contours at the highest morph levels to nonlinear processes and should rely on statistical tests for discovery. As we explain below, statistical tests fail to detect nonlinearity in the integration of features.

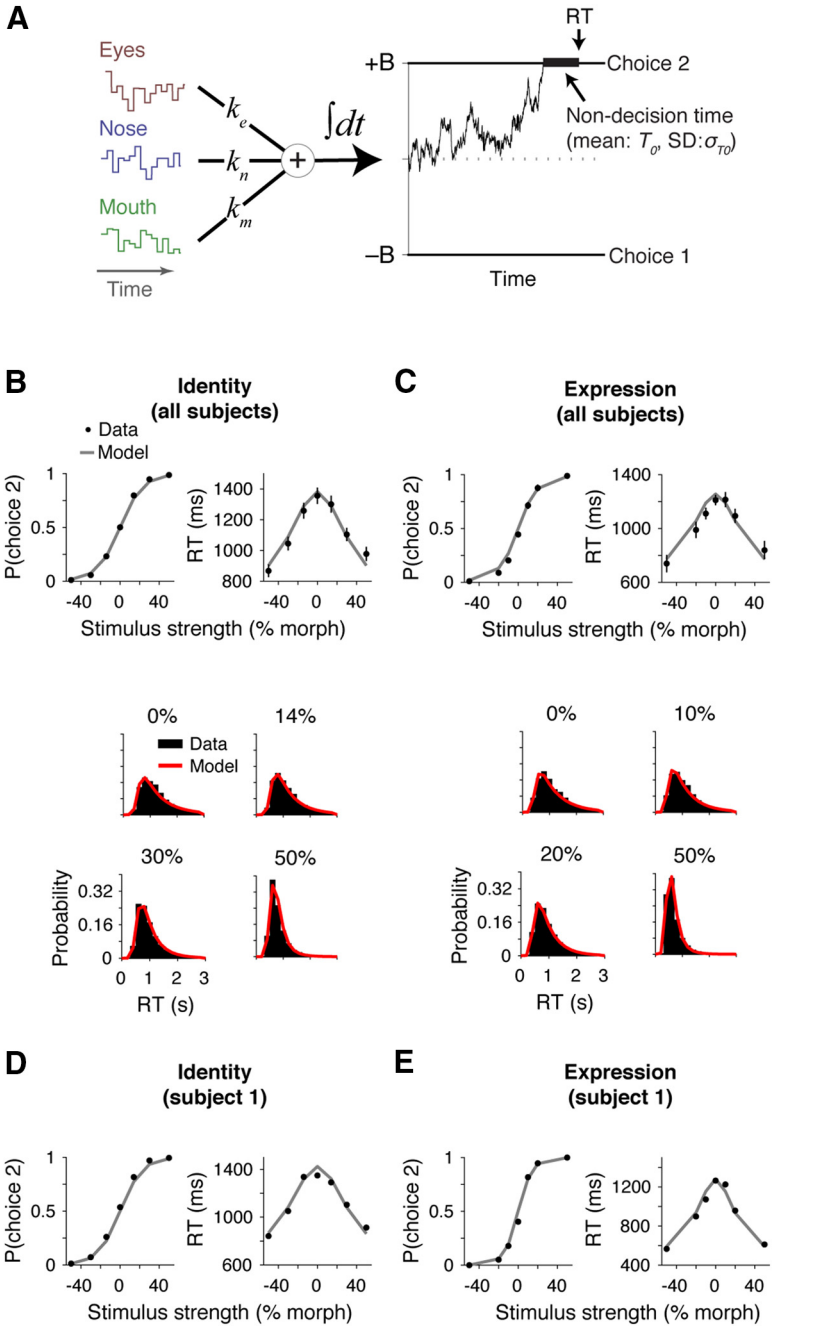


Figure 5. A model that linearly integrates sensory evidence over space and time accounted for choice and reaction time. ***A***, Multifeature drift diffusion model. The model linearly integrates the morph levels of the three informative features with static spatial sensitivities (k_e, k_n, k_m) to create the momentary evidence, which is then integrated over time to create a decision variable. The integration process continues until the decision variable reaches one of the two decision bounds, corresponding to choices 1 and 2. Reaction time equals the time to reach the bound (decision time) plus a nondecision time that reflects the aggregate of sensory and motor delays. ***B***, ***C***, Model fit to choices, mean reaction times, and reaction time distributions in the identity (***B***) and expression (***C***) categorization tasks. Data points (dots) are the same as those in Figure 2*A–D*. The reaction time distribution for each stimulus strength was generated based on all trials with the absolute strength matching the plot title. Black bars are the data, and red lines are model fits. ***D***, ***E***, The model accurately fits all single subject's data. An example subject is shown.

To quantify the contributions of linear and nonlinear factors, we performed a logistic regression on the choices using both linear and nonlinear multiplicative combinations of the feature strengths (Eq. 4). The model accurately fit to the contour lines in Figure 3, *A* and *B* (thick pale lines; identity task, $R^2 = 0.998$; expression task, $R^2 = 0.999$; the thickness of the lines reflects 2 SEM of the logistic parameters). The model coefficients (Fig. 3*C*,*D*) show significant positive sensitivities for the linear effects of all facial features in the identity task (eyes, $t_{(11)} = 10.5$, $p = 4.3 \times 10^{-7}$; nose, $t_{(11)} = 4.6$, $p = 8.2 \times 10^{-4}$; mouth, $t_{(11)} = 4.8$, $p = 5.2 \times 10^{-4}$, two-tailed t test) and in the expression task (eyes, $t_{(6)} = 4.7$, $p = 0.0032$;

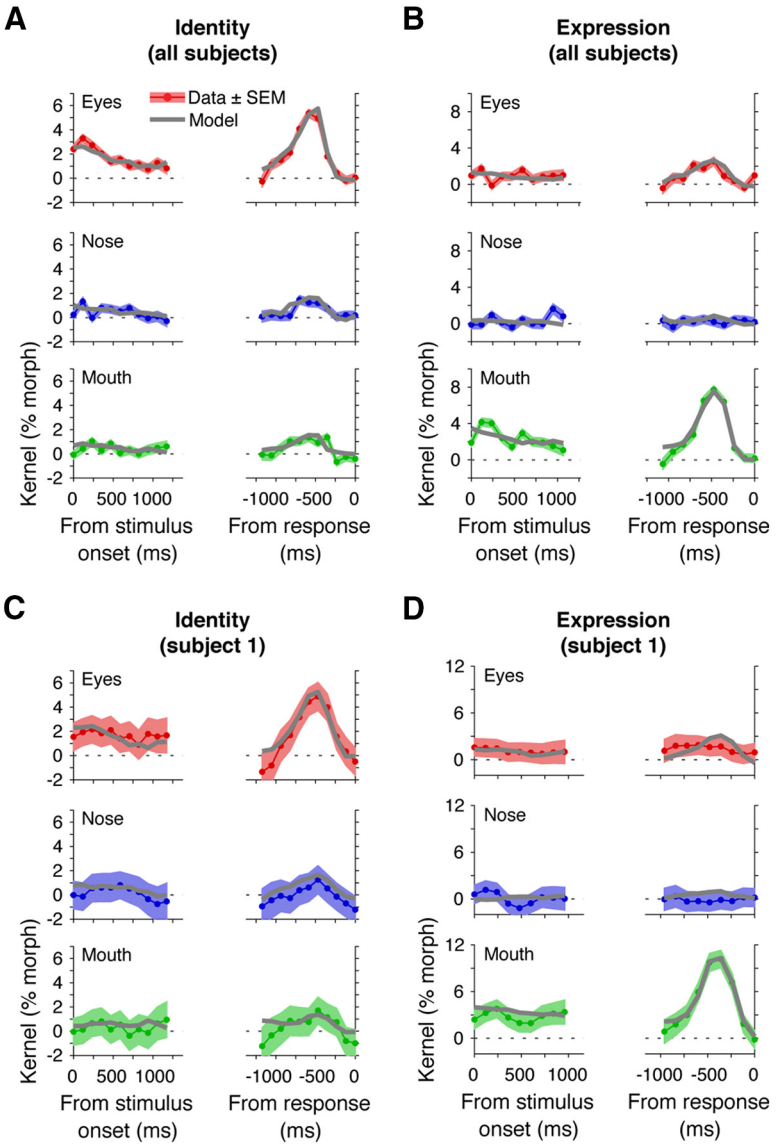


Figure 6. The linear spatiotemporal integration model accurately accounts for the dynamics of psychophysical kernels. ***A***, ***B***, Dynamics of psychophysical kernels averaged across subjects. Shading indicates SEM. Subjects' decisions are most strongly influenced by the eye information in the identity task and the mouth information in the expression task, evidenced by the differential amplitudes of individual feature psychophysical kernels. Gray lines are model fits. Note that feature sensitivities are fixed in the model and the non-stationary kernels do not indicate changes of sensitivity over time (Okazawa et al., 2018). ***C***, ***D***, The model accurately accounts for single subject's kernels. An example subject (the same as Fig. 5*D*,*E*) is shown.

nose, $t_{(6)} = 3.4$, $p = 0.015$; mouth, $t_{(6)} = 9.2$, $p = 9.5 \times 10^{-5}$), but no significant effect for nonlinear terms ($p > 0.27$ for all multiplicative terms in both tasks). These results were largely consistent for individual subjects too (11 of 12 subjects in identity and 7 of 7 in expression task showed no significant improvement in fitting performance by adding nonlinear terms, $p > 0.05$; likelihood ratio test, Bonferroni corrected across subjects). Overall, linear integration provides an accurate and parsimonious account of how features were combined over space for face categorization.

Linearity of the mapping between stimulus strength and subjective evidence

Quantitative prediction of behavior requires understanding the mapping between the stimulus strength as defined by the experimenter (morph level in our experiment) and the evidence conferred by the stimulus for the subject's decision. The parallel linear contours in Figure 3 demonstrate that the strength of one informative feature can be traded for another informative feature to maintain the same choice probability. They further show that this trade-off is largely stable across the stimulus space, strongly suggesting a linear mapping between morph levels and inferred evidence.

To formally test this hypothesis, we quantified the relationship between feature strengths and the effects of features on choice by estimating subjective evidence in log-odds units. Following the methods developed by Yang and Shadlen (2007), we split the feature strengths (% morph) of each stimulus frame into 10 levels and performed a logistic regression to explain subjects' choices based on the number of occurrences of different feature morph levels in a trial. The resulting regression coefficients correspond to the change of the log odds of choice furnished by a feature morph level. For both the identity and expression morphs, the stimulus strength mapped linearly onto subjective evidence (Fig. 4; identity task, $R^2 = 0.94$; expression task, $R^2 = 0.96$), with the exception of the highest stimulus strengths, which exerted slightly larger effects on choice than expected from a linear model. The linearity for a wide range of morph levels—especially for the middle range in which subjects frequently chose both targets—permits us to approximate the total evidence conferred by a stimulus as a weighted sum of the morph levels of the informative features.

Temporal integration mechanisms

The linearity of spatial integration significantly simplifies our approach to investigate integration of sensory evidence over time. We adopted a quantitative-model-based approach by testing a variety of models that have the same linear spatial integration process but differ in ways that use stimulus information over time. We leveraged stimulus fluctuations within and across trials to identify the mechanisms that shaped the behavior. We further validated these models by comparing predicted psychophysical kernels with the empirical ones.

In our main model, the momentary evidence from each stimulus frame is linearly integrated over time (Fig. 5*A*). The momentary evidence from a stimulus frame is a linear weighted sum of the morph levels of informative features in the stimulus, compatible with linear spatial integration shown in the previous sections. The model assumes that sensitivities for these informative features (k_e, k_n, k_m) are fixed within a trial. However, because the stimulus is dynamic and stochastic, the rate of increase of accumulated evidence varies over time. The decision-making process is noisy, with the majority of noise originating from the

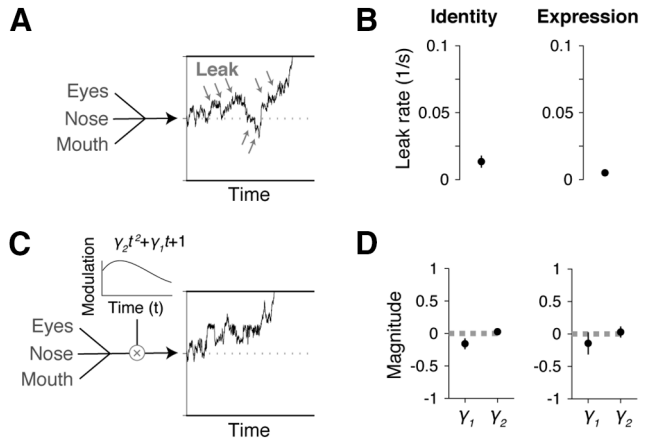


Figure 7. Spatiotemporal integration has static sensitivity to features and minimal forgetting (leak). We used two extensions of the multifeature drift diffusion model to test for leaky integration and modulation of feature sensitivities over time. ***A***, Schematic of the leaky integration model. An exponential leak term was added to the temporal integration process to examine potential loss of information over time (Eq. 9). The leak pushes the decision variable toward zero (gray arrows), reducing the effect of earlier evidence. ***B***, Fitting the leaky integration model to behavioral data revealed leak rates close to zero for both the identity and expression categorization tasks. ***C***, Schematic of the model with dynamic modulation of feature sensitivities. We added a second-order polynomial modulation function to investigate potential temporally nonuniform influence of stimulus features on choice within a trial (Eq. 10). The function allows for a variety of temporal patterns, including ascending, descending, and nonmonotonic changes. ***D***, Fitting the model revealed no substantial change in sensitivities over time, with both the linear and quadratic terms of the modulation function close to zero. Error bars indicate SEM across subjects.

stimulus representation in sensory cortices and inference of momentary evidence from these representations (Brunton et al., 2013; Drugowitsch et al., 2016; Waskom and Kiani, 2018). Because of the noise, the decision-making process resembles a diffusion process with variable drift over time. The process stops when the decision variable (accumulated noisy evidence) reaches one of the two bounds, corresponding to the two competing choices in the task. The bound that is reached dictates the choice, and the reaction time equals the time to bound (decision time) plus a nondecision time composed of sensory and motor delays (Smith and Vickers, 1988; Link, 1992; Ratcliff and Rouder, 2000; Gold and Shadlen, 2007). For each stimulus sequence, we calculated the probability of different choices and expected distribution of reaction times, adjusting model parameters to best match the model with the observed behavior (maximum likelihood fitting of the joint distribution of choice and RT; see Materials and Methods). The model accurately explained subjects' choices (identity, $R^2 = 0.99 \pm 0.002$; expression: $R^2 = 0.98 \pm 0.005$, mean $\pm$ SEM across subjects) and mean reaction times (identity, $R^2 = 0.86 \pm 0.05$; expression, $R^2 = 0.81 \pm 0.05$) as well as the distributions of reaction times (Fig. 5*B*,*C*). The accurate match between the data and the model was also evident within individual subjects (Fig. 5*D*,*E*).

The same model also quantitatively explains the psychophysical kernels (Fig. 6; identity task, $R^2 = 0.86$; expression task, $R^2 = 0.84$). The observed kernels showed characteristic temporal dynamics in addition to the inhomogeneity of amplitudes across features, as described earlier (Fig. 2*E*,*F*). The temporal dynamics are explained by decision bounds and nondecision time in the model (Okazawa et al., 2018). When aligned to stimulus onset, the kernels decreased over time. This decline happens in the model because nondecision time creates a temporal gap between bound crossing and the report of the decision, making stimuli immediately before the report inconsequential for the decision. When aligned to the saccadic response, the kernels peaked several hundred milliseconds before the saccade. This peak emerges in the model because stopping is conditional on a stimulus fluctuation that takes the decision variable beyond the bound, whereas the drop near the response time happens again because of the nondecision time. Critically, the model assumptions about static sensitivity and linear integration matched the observed kernels. Further, the inequality of kernel amplitudes across facial features and tasks were adequately captured by the different sensitivity parameters for individual features (k_e, k_n, k_m) in the model.

To further test properties of temporal integration, we challenged our model with two plausible extensions (Fig. 7). First, integration may be imperfect, and early information can be gradually lost over time (Usher and McClelland, 2001; Bogacz et al., 2006). Such a leaky integration process can be modeled by incorporating an exponential leak rate in the integration process (Fig. 7*A*). When this leak rate becomes infinitely large, the model reduces to a memory-less process that commits to a choice if the momentary sensory evidence exceeds a decision bound, that is, extrema detection (Waskom and Kiani, 2018; Stine et al., 2020). To examine these alternatives, we fit the leaky integration model to the behavioral data. Although the leak rate is difficult to assess in typical perceptual tasks (Stine et al., 2020), our temporally fluctuating stimuli provide a strong constraint on the range of the leak rate that matches behavioral data because increased leak rates lead to lower contribution of earlier stimulus fluctuations to choice. We found that although the fitted leak rate was statistically greater than zero (Fig. 7*B*; identity task, $t_{(11)} = 3.01$, $p = 0.012$; expression task, $t_{(6)} = 2.99$, $p = 0.024$), it was consistently small across subjects (identity task, mean $\pm$ SEM across subjects, $0.013 \pm 0.004 s^{-1}$; expression task, $0.005 \pm 0.002 s^{-1}$). These leak rates correspond to integration time constants larger than 100 s, which is much longer than the duration of each trial (~1 s), supporting near-perfect integration over time.

The second extension allows time-varying sensitivity to sensory evidence within a trial (Levi et al., 2018), as opposed to the constant sensitivity assumed in our main model. To capture a wide variety of plausible temporal dynamics, we added linear and quadratic temporal modulations of drift rate over time to the model (Fig. 7*C*; Eq. 10). However, the modulation parameters were quite close to zero (Fig. 7*D*; identity task: $\gamma_1 = -0.16 \pm 0.084$, $t_{(11)} = -1.88$, $p = 0.087$; $\gamma_2 = 0.027 \pm 0.032$, $t_{(11)} = 0.85$, $p = 0.41$; expression task: $\gamma_1 = -0.15 \pm 0.17$, $t_{(6)} = -0.86$, $p = 0.43$; $\gamma_2 = 0.029 \pm 0.085$, $t_{(6)} = 0.34$, $p = 0.75$, two-tailed t test), suggesting a lack of substantial temporal dynamics. Note, however, that the models with slow modulations of sensitivity cannot capture the very fast modulations in the observed psychophysical kernels. Although the fits might improve by allowing fast modulations of sensitivity (5–10 Hz), we are unaware of sensory or decision mechanisms that can create such fast fluctuations. These fluctuations likely arise from noise because of finite samples in our dataset. Overall, the temporal properties of the decision-making process are consistent with linear multifeature integration with largely static sensitivities.

Testing the sequence of spatiotemporal integration

In the models above, temporal integration operates on the momentary evidence generated from the spatial integration of features of each stimulus frame. But is it necessary for spatial integration to precede temporal integration? Although our data-driven analyses above suggest that subjects combined

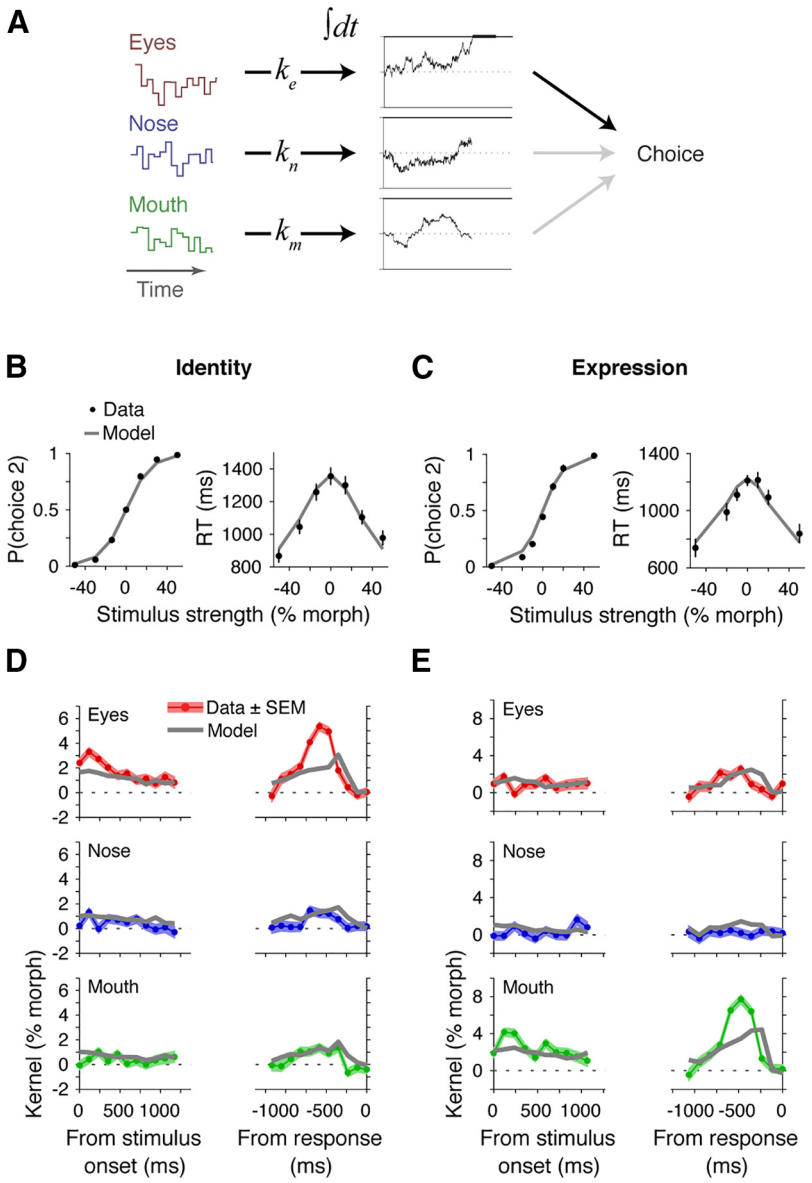


Figure 8. Models with late spatial integration across features fail to explain the experimental data. ***A***, Schematic of a parallel integration model in which single features are independently integrated over time to a decision bound. The first feature that reaches the bound dictates the choice (see Fig. 9 for alternative decision rules). ***B***, ***C***, Model fits to psychometric and chronometric functions. Conventions are the same as in Figure 5 *B* and *C*. ***D***, ***E***, The model fails to explain psychophysical kernels. Notable discrepancies with the data are visible in the kernels for eyes in the identity task and the kernels for mouth in the expression task. Conventions are the same as in Figure 6.

information across facial features (Figs. 3, 4), it might be plausible that spatial integration follows the temporal integration process instead of preceding it. Specifically, the evidence conferred by each informative facial feature may be independently integrated over time, and then a decision may be rendered based on the collective outcome of the three feature-wise integration processes (i.e., spatial integration following temporal integration). A variety of spatial pooling rules may be used in such a model. A choice can be determined by the first feature integrator that reaches the bound (Fig. 8*A*), by the majority of feature integrators (Fig. 9A), or by the sum of the decision variables of the integrators after the first bound crossing (Fig. 9*B*). In all of these model variants, the choice is shaped by multiple features because of the stochasticity of the stimuli and noise (Otto and Mamassian, 2012). For example, the eyes integrator would dictate the choice in many trials of the identity categorization task, but the other feature integrators would also have a smaller but tangible effect, compatible with the differential contribution of features to choice, as shown in the previous sections (Fig. 2*E*,*F*). Are such parallel integration models compatible with the empirical data?

Figure 8 demonstrates that models with late spatial integration fail to explain the behavior. Although these models could fit the psychometric and chronometric functions (Fig. 8*B*,*C*), they underperformed our main model (model log-likelihood difference for the joint distribution of choice and RT, −475.5 in the identity task and −307.6 in the expression task). Moreover, and critically, they did not replicate the subjects' psychophysical kernels (Fig. 8*D*,*E*; identity task, $R^2 = 0.46$; expression task, $R^2 = 0.51$). They systematically underestimated saccade-aligned kernel amplitudes for the dominant feature of each task (eyes for identity categorization and mouth for expression categorization). Further, the predicted model kernels peaked closer to the saccade onset than the empirical kernels. Because the psychophysical kernel amplitude is inversely proportional to the decision bound (Okazawa et al., 2018), the lower amplitude of these model kernels suggests that the model overestimated the decision bound, which necessitated a shorter nondecision time to compensate for the elongated decision times caused by the higher bounds. These shorter nondecision times pushed model kernel peaks closer to the saccade onset.

In general, late spatial integration causes a lower signal-to-noise ratio and is therefore more prone to wrong choices because it ignores part of the available sensory information by terminating the decision-making process based on only one feature or by suboptimally pooling across spatial features after the termination (Fig. 9, test of different spatial pooling rules). To match subjects' high performance, these models would therefore have to alter the speed accuracy trade-off by pushing the decision bound higher than those used by the subjects. However, this change leads to qualitative distortions in the psychophysical kernels. Our approach to augment standard choice and RT analyses with psychophysical reverse correlations was key to identify these qualitative differences (Okazawa et al., 2018), which can be used to reliably distinguish models with different orders of spatial and temporal integration.

What underlies differential contribution of facial features to choice: visual discriminability or decision weight?

The psychophysical kernels and decision-making models in the previous sections indicated that subjects' choices were differentially

sensitive to fluctuations of the three informative features in each categorization task (Figs. 2*E*,F; 3*C*,*D*; 4) and across tasks (drift diffusion model sensitivity for features depicted in Fig. 10*E*; $F_{(2,51)} = 47.4$, $p = 2.3 \times 10^{-12}$, two-way ANOVA interaction between features and tasks). However, as explained earlier, a higher overall sensitivity to a feature could arise from better visual discriminability of changes in the feature or a higher weight applied by the decision-making process to the feature (Fig. 10*A*). Both factors are likely present in our task. Task-dependent changes of feature sensitivities support the existence of flexible decision weights. Differential visual discriminability is a likely contributor too because of distinct facial features across faces in the identity task or expressions in the expression task. To determine the exact contribution of visual discriminability and decision weights to the overall sensitivity, we measured the discrimination performance of the same subjects for each facial feature using two tasks —odd-one-out discrimination (Fig. 10*B*) and categorization of single facial features (Fig. 11*A*).

In the odd-one-out task, subjects viewed three consecutive images of a facial feature (eyes, nose, or mouth) with different stimulus strengths and chose the one that was perceptually distinct from the other two (Fig. 10*B*). Subjects successfully identified the morph level that was distinct from the other two and had higher choice accuracy when the morph level differences were larger (Fig. 10*C*). However, the improvement of accuracy as a function of morph level difference was not identical across features. The rate of increase (slope of psychometric functions) was higher for the eyes of the identity-task stimuli and higher for the mouth of the expression-task stimuli, suggesting that the most sensitive features in those tasks were most discriminable too. We used a model based on signal detection theory (Maloney and Yang, 2003) to fit the psychometric functions and retrieve the effective representational noise (σ_f) for each facial feature (Fig. 10*D*; see Materials and Methods). As expected from the psychometric functions (Fig. 10*C*), visual discriminability, defined as the inverse of the representational noise in the odd-one-out task, was slightly higher for the eyes of the identity-task stimuli and for the mouth of the expression-task stimuli (Fig. 10*F*; identity task, $F_{(2,16)} = 7.4$, $p = 0.0054$; expression task, $F_{(2,4)} = 22.7$, $p = 0.0066$, repeated measures ANOVA).

Similar results were also obtained in the single-feature categorization tasks, where subjects performed categorizations similar to the main task while viewing only one facial feature (Fig. 11*A*). We derived the model sensitivity for each facial feature by fitting a drift diffusion model to the subjects' choices and RTs (Fig. 11*B*). Because subjects discriminated a single feature in this task, differential weighting of features could not play a role in shaping their behavior, and the model sensitivity for each feature was proportional to the feature discriminability. The order of feature discriminability was similar to that from the odd-one-out task, with eyes showing more discriminability for the stimuli of the identity task (Fig. 11*C*).

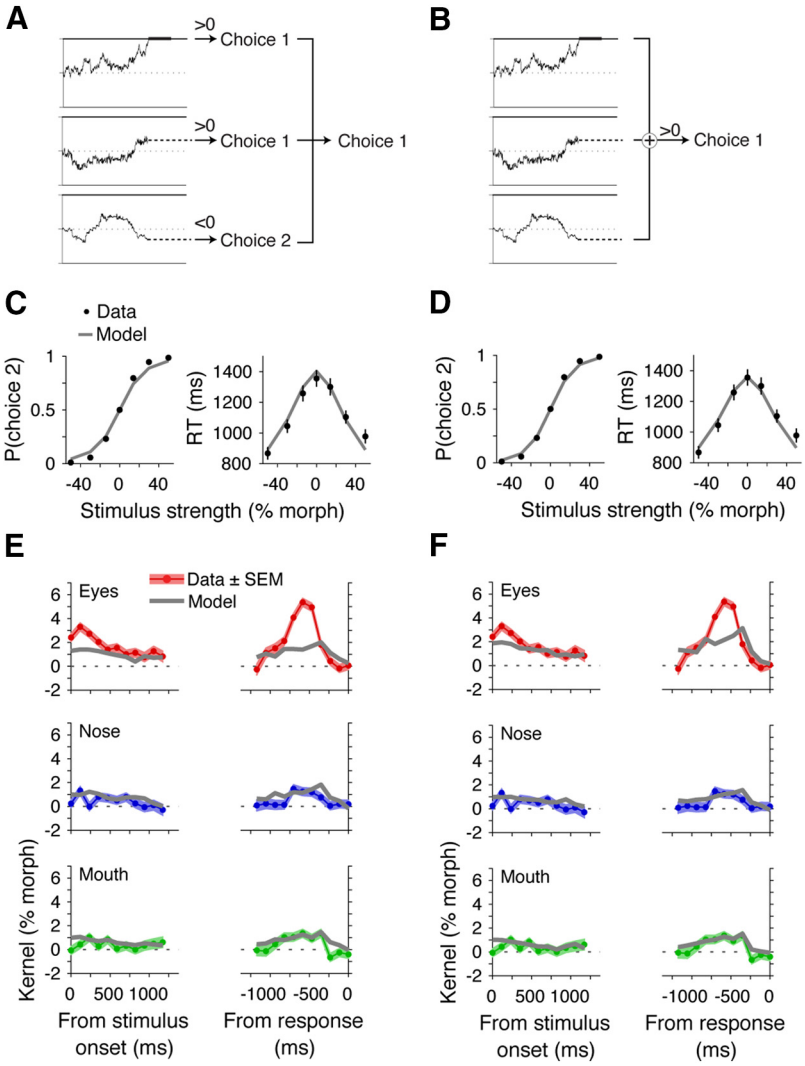


Figure 9. The mismatch between data and late spatial integration models persists with different decision rules. ***A***, A parallel integration model that independently accumulates evidence of the three facial features and commits to a choice favored by the majority of the integrators. When one of the three integrators reaches a bound, the model determines the preferred choice of each integrator based on the sign of the decision variable of the integrator. The model then chooses the option supported by the majority of the integrators (i.e., two or more). ***B***, Another variant of the parallel integration model that renders a decision based on the summed decision variables of the three integrators. When one integrator reaches a bound, the decision variables of the three accumulators are added, and a decision is made based on the sign of the total evidence. ***C***, ***D***, Model fits to psychometric and chronometric functions. The plots show the results for the identity task. ***E***, ***F***, Both models fail to account for the dynamics of psychophysical kernels. Similar results were obtained for the expression task.

Although the results of both tasks support that visual discriminability was nonuniform across facial features, this contrast was less pronounced than that of the model sensitivities in the main task (Fig. 10*E*,*F*). Consequently, dividing the model sensitivities by the discriminability revealed residual differences reflecting nonuniform decision weights across features (Fig. 10*G*; $F_{(2,30)} = 6.1$, $p = 0.0059$, two-way ANOVA, main effect of features) and between the tasks ($F_{(2,30)} = 10.9$, $p = 2.8 \times 10^{-4}$, two-way ANOVA, interaction between features and tasks). In other words, context-dependent decision weights play a significant role in the differential contributions of facial features to decisions. Furthermore, these weights suggest that subjects rely more on more informative (less noisy) features. In fact, the decision weights were positively correlated with visual discriminability (Fig. 10*H*; $R = 0.744$, $p = 2.0 \times 10^{-7}$), akin to an

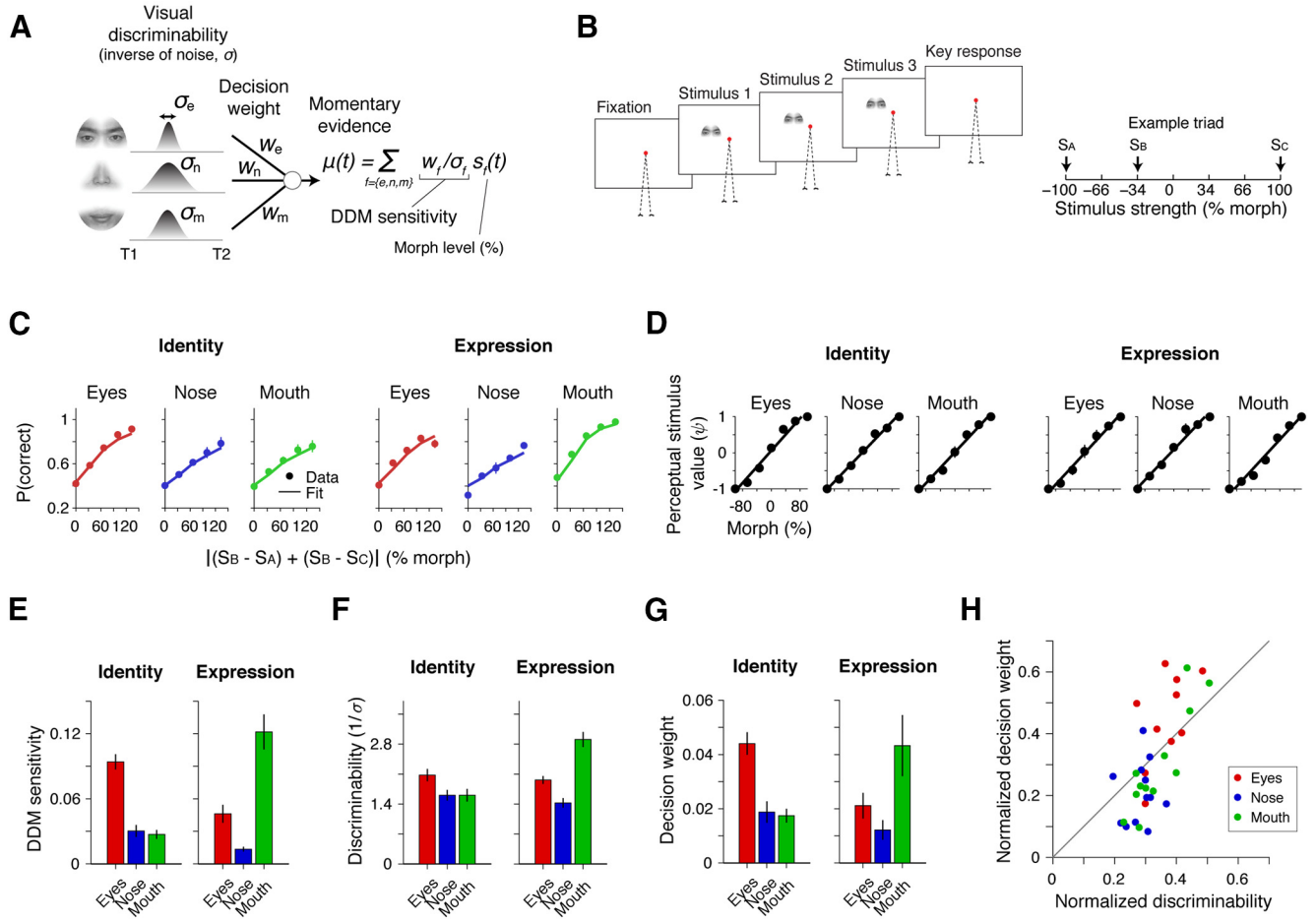


Figure 10. Differential sensitivity of decisions to facial features arises from a combination of visual discriminability and decision weights. ***A***, Schematic of the factors that shape differential sensitivities to the informative features in the drift diffusion model (DDM) depicted in Figure 5*A*. Feature sensitivity in a task arises from the following factors: (1) visual discriminability of different morph levels of the feature and (2) the weight that the decision-making process attributes to the feature. These factors are distinct as visual discriminability arises from the precision of sensory representations, whereas decision weights are flexibly set to achieve a particular behavioral goal. We define visual discriminability of a feature as the inverse of representational noise in units of %morph (σ_f, where f could be e, n, or m for eyes, nose, and mouth, respectively), and the overall sensitivity to a feature as visual discriminability ($1/\sigma_f$) multiplied by the decision weight (w_f). To determine the relative contribution of these two factors, one needs to measure the visual discriminability of facial features. ***B***, The design of an odd-one-out discrimination task to measure the visual discriminability of individual features. In each trial, subjects viewed a sequence of three images of the same feature and reported the one that was perceptually most distinct from the other two. The three images had distinct morph levels (S_A, S_B, S_C, sorted in ascending order). ***C***, Subjects' accuracy as a function of the distinctness of the correct stimulus from the other two. Distinctness was quantified as $|(S_B - S_A) + (S_B - S_C)|$. The lines are the fits of an ideal observer model (see Materials and Methods). The accuracy for 0% distinctness was slightly larger than a random choice (33%) because subjects were slightly less likely to choose the middle morph level (S_B). ***D***, We explained the subjects' responses based on the representational noise of a feature (σ) and the relative distance between individual pairs of stimuli in a perceptual space (ψ_i and ψ_j for stimuli i and j; Eq. 11). The ψ for intermediate morph levels and σ were estimated from data using a maximum likelihood fit. The ψ changes largely linearly with morph levels, ensuring that one can use the inverse of the representational noise ($1/\sigma$) as a metric for the visual discriminability of each facial feature. Data points are the estimated ψ, and lines are the best least squares fits. ***E***, The sensitivity parameters of the multifeature drift diffusion model (Fig. 5*A*) for the informative features in each task. ***F***, Visual discriminability ($1/\sigma$) estimated using the ideal observer model. Facial features have different discriminability with the eyes slightly more discriminable for the faces used in the identity categorization task and the mouth more discriminable in the expression categorization task. However, these differences in discriminability across features are less pronounced than those in the overall sensitivity (***E***). ***G***, Decision weights (w in ***A***) calculated by dividing the overall sensitivity by the visual discriminability of each feature. ***H***, Positive correlation between the visual discriminability (***F***) and the decision weights (***G***) of features is consistent with optimal cue combination. Each dot corresponds to the values of one facial feature of one subject. The plot aggregates data from both the identity and expression tasks. Both the discriminability and decision weights are normalized within subjects (the sum across all features is fixed to 1) to account for the variability of absolute discriminability and weights across subjects.

optimal cue integration process (Ernst and Banks, 2002; Oruç et al., 2003; Drugowitsch et al., 2014). Together, the decision-making process in face categorization involves context-dependent adjustment of decision weights that improves behavioral performance.

Discussion

Successful categorization or identification of objects depends on elaborate sensory and decision-making processes that transmit and use sensory information to implement goal-directed behavior. The properties of the decision-making process remain underexplored for object vision. Existing models commonly assume instantaneous decoding mechanisms based on linear readout of population responses of sensory neurons (Hung et al., 2005; Majaj et al., 2015; Rajalingham et al., 2015; Chang and Tsao, 2017), but they are unable to account for aspects of behavior that are based on deliberation on temporally extended visual information common in our daily environments. By extending a quantitative framework developed for studying simpler perceptual decision (Ratcliff and Rouder, 1998; Palmer et al., 2005; Gold and Shadlen, 2007; O'Connell et al., 2012; Waskom et al.,

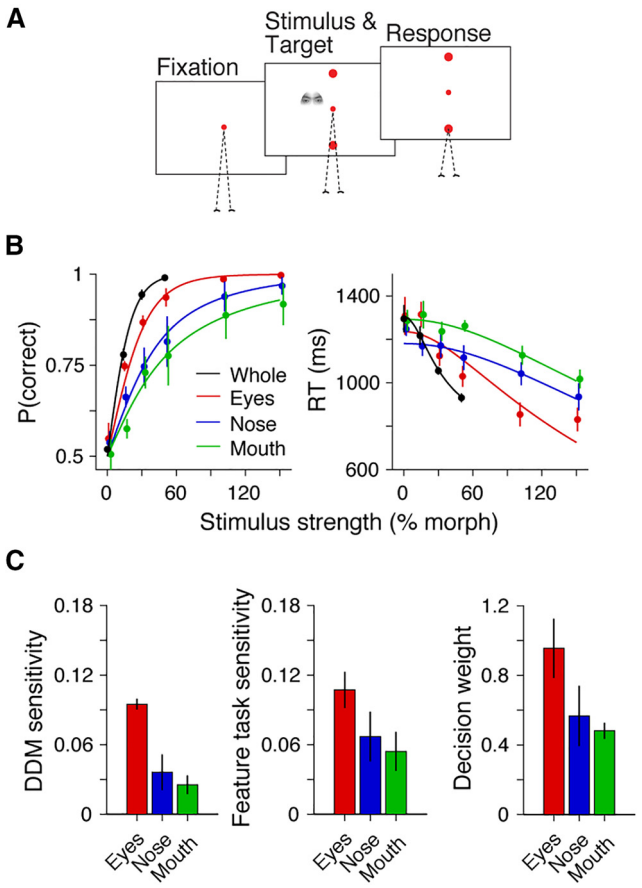


Figure 11. Visual discriminability assessed using a single feature categorization task supports nonuniform decision weights in the main task. ***A***, To confirm the results of the odd-one-out task, we also performed a single-feature categorization task. The subjects categorized the facial identities as in the identity task but based their decisions only on one facial feature shown on each trial. ***B***, Psychometric and chronometric functions for each facial feature. As a comparison, the same subjects' performance in the identity task is shown in black. The lines are the fits of a drift diffusion model for each facial feature. ***C***, Comparison of the model feature sensitivities in the main task (left) and in the single feature categorization task (middle). Dividing the feature sensitivities of the two tasks yields the decision weight for each feature. The results support unequal weighting over features (right), consistent with the results of the odd-one-out task (Fig. 10*E–G*).

2019), we establish an experimental and modeling approach that quantitatively links sensory inputs and behavioral responses during face categorization. We show that human face categorization constitutes spatiotemporal evidence integration processes. A spatial integration process aggregates stimulus information into momentary evidence, which is then integrated over time by a temporal integration process. The temporal integration is largely linear and because of long time constants has minimal or no loss of information over time. The spatial integration is also linear and accommodates flexible behavior across tasks by adjusting the weights applied to visual features. These weights remain stable over time in our task, providing no indication that the construction of momentary evidence or the informativeness changes with stimulus viewing time.

Our approach bridges past studies on object recognition and perceptual decision-making by formulating face recognition as a process that integrates sensory evidence over space and time. Past research on object recognition focused largely on feedforward visual processing and instantaneous readout of the visual representations, leaving a conceptual gap for understanding the temporally extended processes that underlie perception and action planning based on visual object information. Several studies have attempted to fill this gap by using noisy object stimuli (Heekeren et al., 2004; Philiastides and Sajda, 2006; Ploran et al., 2007; Philiastides et al., 2014; Heidari-Gorji et al., 2021) or sequential presentation of object features (Ploran et al., 2007; Jack et al., 2014). However, the stimulus manipulations in these studies did not allow a comprehensive exploration of both spatial and temporal processes. They either created a one-dimensional stimulus axis that eroded the possibility to study spatial integration across features or created temporal sequences that eroded the possibility to study temporal integration jointly with spatial integration. Our success hinges on a novel stimulus design, namely, independent parametric morphing of individual facial features and subliminal spatiotemporal feature fluctuations within trials. Independent feature fluctuations were key to characterize the integration processes, and the subliminal sensory fluctuations ensured that our stimulus manipulations did not alter subjects' decision strategy, addressing a fundamental challenge (Murray and Gold, 2004) to alternative methods (e.g., covering face parts; Gosselin and Schyns, 2001; Schyns et al., 2002; but see Gosselin and Schyns, 2004).

We used three behavioral measures—choice, reaction time, and psychophysical reverse correlation—to assess the mechanisms underlying the behavior. Some key features of the decision-making process cannot be readily inferred solely from choice and reaction time, for example, the time constant of the integration process (Ditterich, 2006; Stine et al., 2020). However, the inclusion of psychophysical kernels provides a more powerful three-pronged approach (Okazawa et al., 2018) that enabled us to establish differential sensitivities for informative features (Fig. 2*E*,F), linearity of spatial integration (Fig. 3), long time constants (minimum information loss) for temporal integration (Fig. 7*B*), static feature sensitivities (Fig. 7*D*), and failure of late spatial integration in the parallel feature integration models (Figs. 8, 9). The precise agreement of psychophysical kernels between model and data (Fig. 6) reinforces our conclusion that face categorization arises from linear spatiotemporal integration of visual evidence.

Face perception is often construed as a holistic process because breaking the configuration of face images, for example, removing parts (Tanaka and Farah, 1993), shifting parts (Young et al., 1987), or inverting images (Yin, 1969), reduces performance for face discrimination (Taubert et al., 2011), categorization (Young et al., 1987), or recognition (Tanaka and Farah, 1993). However, the mechanistic underpinnings of these phenomena remain elusive (Richler et al., 2012). The linear spatial integration mechanism has the potential to provide mechanistic explanations for some of these holistic effects. For example, changes in the configuration of facial features could reduce visual discriminability of facial features (Murphy and Cook, 2017), disrupt spatial integration (Gold et al., 2012; Witthoft et al., 2016), or cause suboptimal weighting of informative features (Sekuler et al., 2004). Holistic effects can also be manifested as impairment in facial part recognition when placed together with other uninformative facial parts (composite face effect; Young et al., 1987). This might arise because face stimuli automatically trigger spatial integration that combines information from irrelevant parts. Our approach offers a quantitative path to test these possibilities using a unified modeling framework—a fruitful direction to pursue in the future.

The linearity of spatial integration over facial features has been a source of controversy in the past (Gold et al., 2012; Gold, 2014; Shen and Palmeri, 2015). The controversy partly stems from the ambiguity in what visual information contributes to face recognition. Some suggest that local shape information of facial parts accounts for holistic face processing (McKone and Yovel, 2009), whereas others suggest that configural information, such as distances between facial features, gives rise to nonlinearities (Shen and Palmeri, 2015) and holistic properties (Le Grand et al., 2001; Maurer et al., 2002). Our study does not directly address this question because feature locations in our stimuli were kept largely constant to facilitate morphing between faces. However, our approach can be generalized to include configural information and systematically tease apart spatial integration over feature contents from integration over the relative configuration of features. An ideal decision-making process would treat configural information similar to content information by linearly integrating independent pieces of information. Although our current results strongly suggest linear integration over feature contents, we remain open to emergent nonlinearities for configural information.

Another key finding in our experiments is flexible, task-dependent decision weights for informative features (Fig. 10). Past studies demonstrated the preferential use of more informative features over others during face and object categorization (Schyns et al., 2002; Sigala and Logothetis, 2002; De Baene et al., 2008). But it was not entirely clear whether and by how much subjects' enhanced sensitivity stemmed from visual discriminability of features or decision weights. We have shown that the differential model sensitivity for facial features in our tasks could not be fully explained by inhomogeneity of visual discriminability across features, thus confirming flexible decision weights for facial features. Importantly, the weights were proportional to the visual discriminability of features in each task (Fig. 10*H*), consistent with the idea of optimal cue integration that explains multisensory integration behavior (Ernst and Banks, 2002; Oruç et al., 2003; Drugowitsch et al., 2014). Our observation suggests that face recognition is compatible with Bayesian computations in cue combination paradigms (Gold et al., 2012; Fetsch et al., 2013). It is an important future direction to test whether the recognition of other object categories also conforms to such optimal computations (Kersten et al., 2004). Moreover, neural responses to object stimuli can dynamically change because of adaptation or expectation (Kaliukhovich et al., 2013), which can alter both the sensory and decision-making processes (Mather and Sharman, 2015; Witthoft et al., 2018). How decision-making processes adapt to dynamic inputs is another important direction to be explored in the future.

The quantitative characterization of behavior is pivotal for linking computational mechanisms and neural activity as it guides future research on where and how the spatiotemporal integration of sensory evidence is implemented in the brain. Face stimuli evoke activity in a wide network of regions in the temporal cortex with different levels of specialization for processing facial parts, view invariance, facial identity and emotions, as well as social interactions (Freiwald and Tsao, 2010; Freiwald et al., 2016; Sliwa and Freiwald, 2017; Hesse and Tsao, 2020; Hu et al., 2020). Although neural activity in these regions is known to causally alter face recognition behavior (Afraz et al., 2006; Parvizi et al., 2012; Moeller et al., 2017), the exact contribution to the decision-making process remains unresolved. Prevailing theories emphasize the role of these regions in sensory processing, commonly attributing rigid selectivities to the neurons that are invariant to behavioral goals. In these theories, flexible spatiotemporal integration of evidence, as we explain in our model, is left to downstream sensorimotor or association areas commonly implicated in decision-making (Ratcliff et al., 2003; Cisek and Kalaska, 2005; Gold and Shadlen, 2007; Schall, 2019; Okazawa et al., 2021). However, neurons in the inferior temporal cortex show response dynamics that can reflect temporally extended decisions (Akrami et al., 2009), and they may alter selectivity in a task-dependent manner (Koida and Komatsu, 2007; Tajima et al., 2017), challenging a purely sensory role for the inferior temporal neurons and hinting at the potential contribution of these neurons to flexible spatial and temporal integration. Future studies that focus on the interactions between temporal cortex and downstream areas implicated in decision-making will clarify the role of different brain regions. Our experimental framework provides a foundation for studying such interactions by determining the properties of spatiotemporal integration and making quantitative predictions about the underlying neural responses.

References

Afraz SR, Kiani R, Esteky H (2006) Microstimulation of inferotemporal cortex influences face categorization. Nature 442:692–695.

Ahumada AJ (1996) Perceptual classification images from vernier acuity masked by noise. Perception 25:2–2.

Akrami A, Liu Y, Treves A, Jagadeesh B (2009) Converging neuronal activity in inferior temporal cortex during the classification of morphed stimuli. Cereb Cortex 19:760–776.

Barraclough NE, Perrett DI (2011) From single cells to social perception. Philos Trans R Soc Lond B Biol Sci 366:1739–1752.

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD (2006) The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev 113:700–765.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436.

Brunton BW, Botvinick MM, Brody CD (2013) Rats and humans can optimally accumulate evidence for decision-making. Science 340:95–98.

Carlson TA, Ritchie JB, Kriegeskorte N, Durvasula S, Ma J (2014) Reaction time for object categorization is predicted by representational distance. J Cogn Neurosci 26:132–142.

Chang L, Tsao DY (2017) The code for facial identity in the primate brain. Cell 169:1013–1028 e14.

Cisek P, Kalaska JF (2005) Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. Neuron 45:801–814.

De Baene W, Ons B, Wagemans J, Vogels R (2008) Effects of category learning on the stimulus selectivity of macaque inferior temporal neurons. Learn Mem 15:717–727.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341.

Ditterich J (2006) Stochastic models of decisions about motion direction: behavior and physiology. Neural Netw 19:981–1012.

Drugowitsch J, DeAngelis GC, Klier EM, Angelaki DE, Pouget A (2014) Optimal multisensory decision-making in a reaction-time task. eLife 3: e03005.

Drugowitsch J, Wyart V, Devauchelle AD, Koechlin E (2016) Computational precision of mental inference as critical source of human choice suboptimality. Neuron 92:1398–1411.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429–433.

Fetsch CR, DeAngelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429–442.

Freiwald W, Duchaine B, Yovel G (2016) Face processing systems: from neurons to real-world social perception. Annu Rev Neurosci 39:325–346.

Freiwald WA, Tsao DY (2010) Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science 330:845–851.

Gauthier I, Anderson AW, Tarr MJ, Skudlarski P, Gore JC (1997) Levels of categorization in visual recognition studied using functional magnetic resonance imaging. Curr Biol 7:645–651.

Gauthier I, Williams P, Tarr MJ, Tanaka J (1998) Training 'greeble' experts: a framework for studying expert object recognition processes. Vision Res 38:2401–2428.
Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574.
Gold JM (2014) A perceptually completed whole is less than the sum of its parts. Psychol Sci 25:1206–1217.
Gold JM, Mundy PJ, Tjan BS (2012) The perception of a face is no more than the sum of its parts. Psychol Sci 23:427–434.
Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vision Res 41:2261–2271.
Gosselin F, Schyns PG (2004) No troubles with bubbles: a reply to Murray and Gold. Vision Res 44:471–477.
Hanks T, Kiani R, Shadlen MN (2014) A neural mechanism of speed-accuracy tradeoff in macaque area LIP. eLife 3:e02260.
Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG (2004) A general mechanism for perceptual decision-making in the human brain. Nature 431:859–862.
Heidari-Gorji H, Ebrahimpour R, Zabbah S (2021) A temporal hierarchical feedforward model explains both the time and the accuracy of object recognition. Sci Rep 11:5640.
Heitz RP, Schall JD (2012) Neural mechanisms of speed-accuracy tradeoff. Neuron 76:616–628.
Hesse JK, Tsao DY (2020) The macaque face patch system: a turtle's underbelly for the brain. Nat Rev Neurosci 21:695–716.
Hu Y, Baragchizadeh A, O'Toole AJ (2020) Integrating faces and bodies: psychological and neural perspectives on whole person perception. Neurosci Biobehav Rev 112:472–486.
Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310:863–866.
Jack RE, Garrod OG, Schyns PG (2014) Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr Biol 24:187–192.
Kaliukhovich DA, De Baene W, Vogels R (2013) Effect of adaptation on object representation accuracy in macaque inferior temporal cortex. J Cogn Neurosci 25:777–789.
Kampf M, Nachson I, Babkoff H (2002) A serial test of the laterality of familiar face recognition. Brain Cogn 50:35–50.
Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond B Biol Sci 361:2109–2128.
Karlin S, Taylor HE (1981) A second course in stochastic processes. Amsterdam: Elsevier.
Kersten D, Mamassian P, Yuille A (2004) Object perception as bayesian inference. Annu Rev Psychol 55:271–304.
Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324:759–764.
Koida K, Komatsu H (2007) Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. Nat Neurosci 10:108–116.
Kreichman O, Bonneh YS, Gilaie-Dotan S (2020) Investigating face and house discrimination at foveal to parafoveal locations reveals category-specific characteristics. Sci Rep 10:8306.
Le Grand R, Mondloch CJ, Maurer D, Brent HP (2001) Early visual experience and face processing. Nature 410:890.
Levi AJ, Yates JL, Huk AC, Katz LN (2018) Strategic and dynamic temporal weighting for perceptual decisions in humans and macaques. eNeuro 5: ENEURO.0169–18.2018.
Levy I, Hasson U, Avidan G, Hendler T, Malach R (2001) Center-periphery organization of human object areas. Nat Neurosci 4:533–539.
Link SW (1992) The wave theory of difference and similarity. Hillside, NJ: Erlbaum.
Majaj NJ, Hong H, Solomon EA, DiCarlo JJ (2015) Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. J Neurosci 35:13402–13418.
Maloney LT, Yang JN (2003) Maximum likelihood difference scaling. J Vis 3:573–585.
Mather G, Sharman RJ (2015) Decision-level adaptation in motion perception. R Soc Open Sci 2:150418.
Maurer D, Grand RL, Mondloch CJ (2002) The many faces of configural processing. Trends Cogn Sci 6:255–260.
McKone E, Yovel G (2009) Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. Psychon Bull Rev 16:778–797.
Moeller S, Crapse T, Chang L, Tsao DY (2017) The effect of face patch microstimulation on perception of faces and objects. Nat Neurosci 20:743–752.
Murphy J, Cook R (2017) Revealing the mechanisms of human face perception using dynamic apertures. Cognition 169:25–35.
Murray RF, Gold JM (2004) Troubles with bubbles. Vision Res 44:461–470.
O'Connell RG, Dockree PM, Kelly SP (2012) A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nat Neurosci 15:1729–1735.
Okazawa G, Sha L, Purcell BA, Kiani R (2018) Psychophysical reverse correlation reflects both sensory and decision-making processes. Nat Commun 9:3479.
Okazawa G, Hatch CE, Mancoo A, Machens CK, Kiani R (2021) Representational geometry of perceptual decisions in the monkey parietal cortex. Cell 184:3748–3761.e18.
Oruç I, Maloney LT, Landy MS (2003) Weighted linear cue combination with possibly correlated error. Vision Res 43:2451–2468.
Otto TU, Mamassian P (2012) Noise and correlations in parallel perceptual decision making. Curr Biol 22:1391–1396.
Palmer J, Huk AC, Shadlen MN (2005) The effect of stimulus strength on the speed and accuracy of a perceptual decision. J Vis 5:376–404.
Parvizi J, Jacques C, Foster BL, Witthoft N, Withoft N, Rangarajan V, Weiner KS, Grill-Spector K (2012) Electrical stimulation of human fusiform face-selective regions distorts face perception. J Neurosci 32:14915–14920.
Perrodin C, Kayser C, Abel TJ, Logothetis NK, Petkov CI (2015) Who is that? Brain networks and mechanisms for identifying individuals. Trends Cogn Sci 19:783–796.
Philiastides MG, Sajda P (2006) Temporal characterization of the neural correlates of perceptual decision making in the human brain. Cereb Cortex 16:509–518.
Philiastides MG, Heekeren HR, Sajda P (2014) Human scalp potentials reflect a mixture of decision-related signals during perceptual choices. J Neurosci 34:16877–16889.
Ploran EJ, Nelson SM, Velanova K, Donaldson DI, Petersen SE, Wheeler ME (2007) Evidence accumulation and the moment of recognition: dissociating perceptual recognition processes using fMRI. J Neurosci 27:11912–11924.
Rajalingham R, Schmidt K, DiCarlo JJ (2015) Comparison of object recognition behavior in human and monkey. J Neurosci 35:12127–12136.
Ramon M, Caharel S, Rossion B (2011) The speed of recognition of personally familiar faces. Perception 40:437–449.
Ratcliff R, Rouder JN (1998) Modeling response times for two-choice decisions. Psychol Sci 9:347–356.
Ratcliff R, Rouder JN (2000) A diffusion model account of masking in two-choice letter identification. J Exp Psychol Hum Percept Perform 26:127–140.
Ratcliff R, Cherian A, Segraves M (2003) A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. J Neurophysiol 90:1392–1407.
Richler JJ, Palmeri TJ, Gauthier I (2012) Meanings, mechanisms, and measures of holistic processing. Front Psychol 3:553.
Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.
Rossion B (2014) Understanding face perception by means of human electrophysiology. Trends Cogn Sci 18:310–318.
Schall JD (2019) Accumulators, neurons, and response time. Trends Neurosci 42:848–860.
Schyns PG, Bonnar L, Gosselin F (2002) Show me the features! Understanding recognition from the use of visual information. Psychol Sci 13:402–409.
Schyns PG, Petro LS, Smith ML (2007) Dynamics of visual information integration in the brain for categorizing facial expressions. Curr Biol 17:1580–1585.
Sekuler AB, Gaspar CM, Gold JM, Bennett PJ (2004) Inversion leads to quantitative, not qualitative, changes in face processing. Curr Biol 14:391–396.
Shadlen MN, Hanks TD, Churchland AK, Kiani R, Yang T (2006). The speed and accuracy of a simple perceptual decision: A mathematical primer. In

Bayesian brain: probabilistic approaches to neural coding (Doya K, Ishii S, Pouget A, Rao RPN, eds) pp 209–237. Cambridge, MA: MIT.
Shen J, Palmeri TJ (2015) The perception of a face can be greater than the sum of its parts. Psychon Bull Rev 22:710–716.
Sigala N, Logothetis NK (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415:318–320.
Sliwa J, Freiwald WA (2017) A dedicated network for social interaction processing in the primate brain. Science 356:745–749.
Smith PL, Vickers D (1988) The accumulator model of two-choice discrimination. J Math Psychol 32:135–168.
Smith PL, Little DR (2018) Small is beautiful: in defense of the small-N design. Psychon Bull Rev 25:2083–2101.
Stine GM, Zylberberg A, Ditterich J, Shadlen MN (2020) Differentiating between integration and non-integration strategies in perceptual decision making. eLife 9:e55365.
Tajima S, Koida K, Tajima CI, Suzuki H, Aihara K, Komatsu H (2017) Task-dependent recurrent dynamics in visual cortex. eLife 6:e26868.
Tanaka JW, Farah MJ (1993) Parts and wholes in face recognition. Q J Exp Psychol A 46:225–245.
Taubert J, Apthorp D, Aagten-Murphy D, Alais D (2011) The role of holistic processing in face perception: evidence from the face inversion effect. Vision Res 51:1273–1278.
Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522.
Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, Marcus DJ, Westerlund A, Casey BJ, Nelson C (2009) The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res 168:242–249.
Tsao DY, Livingstone MS (2008) Mechanisms of face perception. Annu Rev Neurosci 31:411–437.
Usher M, McClelland JL (2001) The time course of perceptual choice: the leaky, competing accumulator model. Psychol Rev 108:550–592.
Waskom ML, Kiani R (2018) Decision making through integration of sensory evidence at prolonged timescales. Curr Biol 28:3850–3856 e9.
Waskom ML, Okazawa G, Kiani R (2019) Designing and interpreting psychophysical investigations of cognition. Neuron 104:100–112.
Witthoft N, Poltoratski S, Nguyen M, Golarai G, Liberman A, LaRocque KF, Smith ME, Grill-Spector K (2016) Reduced spatial integration in the ventral visual cortex underlies face recognition deficits in developmental prosopagnosia. bioRxiv 051102.
Witthoft N, Sha L, Winawer J, Kiani R (2018) Sensory and decision-making processes underlying perceptual adaptation. J Vis 18:10.
Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624.
Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. Nature 447:1075–1080.
Yin RK (1969) Looking at upside-down faces. J Exp Psychol 81:141–145.
Young AW, Hellawell D, Hay DC (1987) Configurational information in face perception. Perception 16:747–759.
Zhan J, Ince RAA, Rijsbergen N, van Schyns PG (2019) Dynamic construction of reduced representations in the brain for perceptual decision behavior. Curr Biol 29:319–326 e4.